RNPS No. 2142
ISSN 2072-6287
Versión Digital

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada
MINISTERIO DE LA INDUSTRIA BÁSICA

REPORTE TÉCNICO
Reconocimiento de Patrones

SERIE AZUL

Estado actual de los métodos para la recuperación de imágenes de rostros basados en estructuras de indexación

Alain González Nieda y
Heydi Méndez Vázquez

RT_063 **septiembre 2014**

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

RNPS No. 2142
ISSN 2072-6287
Versión Digital

SERIE AZUL

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada
MINISTERIO DE LA INDUSTRIA BÁSICA

REPORTE TÉCNICO
Reconocimiento de Patrones

Estado actual de los métodos para la recuperación de imágenes de rostros basados en estructuras de indexación

Alain González Nieda y
Heydi Méndez Vázquez

RT_063 **septiembre 2014**

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

Tabla de contenido

Estado actual de los métodos para la recuperación de imágenes de rostros basados en estructuras de indexación

Alain González Nieda y Heydi Méndez Vázquez

Equipo de Investigaciones de Biometría, Centro de Aplicaciones de Tecnologías de Avanzada (CENATAV)
La Habana, Cuba
{agonzalez, hmendez}@cenatav.co.cu

RT_063 Serie Azul, CENATAV
Aceptado: 10 de Septiembre de 2014

Resumen. La gran dimensionalidad de las bases de datos empleadas en sistemas de reconocimiento facial y el incremento de imágenes de personas en la web, hacen imprescindible el uso de algoritmos que garanticen escalabilidad y velocidad en la recuperación de imágenes de rostros. En este trabajo se realiza un estudio del estado actual de los principales algoritmos de recuperación de imágenes de rostros basados en contenido (RIRBC) mediante el empleo de estructuras de indexación. Además se analizan y comparan las metodologías propuestas en estos algoritmos con el objetivo de identificar problemas abiertos para continuar la investigación en esta temática.

Palabras clave: recuperación de imágenes de rostros, estructura de indexación, metodología, reconocimiento facial, reducción del espacio de búsqueda.

Abstract. The high dimensionality of the databases used in face recognition systems and the increasing images of people on the web, make essential the use of algorithms to ensure scalability and speed in retrieving images of faces. This paper summarizes the state of the art of the main algorithms used for content-based face image retrieval (CBFIR) by using index structures. In this work the proposed methodologies of these algorithms are also analyzed and compared in order to identify open problems to continue the research in this topic.

Keywords: face image retrieval, indexation structure, methodology, face recognition, search space reduction.

1 Introducción

El constant e crecimiento de información digital almacenada hace que recuperar datos específicos mediante búsquedas sea un problema abierto en la actualidad. Un ejemplo en el caso de las imágenes de rostros, es el incremento de la cantidad de fotografías en Internet, debido a la gran popularidad de servicios de publicación de fotos desde dispositivos digitales y redes sociales, y la necesidad de estos servicios de recuperar fotos específicas [1, 2]. Este problema también está presente en aplicaciones

biométricas que se despliegan en la actualidad, que necesitan identificar una persona por sus rasgos biométricos, entre millones de individuos [3]. Específicamente en la tarea de reconocimiento facial, es imprescindible el acceso a grandes bases de datos de imágenes de rostros. En un sistema de reconocimiento facial común, de una base de datos se extraen y almacenan en un vector los rasgos faciales extraídos de la imagen del rostro correspondiente. El mismo vector de rasgos de un rostro que se desea identificar es comparado con cada uno de los de la base de datos, generalmente en una búsqueda lineal. La similitud entre el rostro a identificar y la imagen de la base de datos se determina mediante la similitud entre sus respectivos vectores de rasgos. Realizar este proceso mediante una búsqueda lineal conlleva un alto costo computacional y un tiempo de ejecución elevado, por lo que no resulta eficiente en grandes bases de datos.

Cuando el número de imágenes a comparar es muy grande, se hace necesario un método de recuperación facial, que permita, de manera eficiente, obtener a partir de la imagen de un rostro un subconjunto con las imágenes más similares que existan en la base de datos o en la web. El reconocimiento facial, a diferencia de la tarea de recuperación, consiste en identificar o verificar la identidad de una persona a partir de la extracción de rasgos de la imagen del rostro[4]. A pesar que los descriptores utilizados en labores de reconocimiento pueden emplearse para la recuperación, no es sencillo aplicarlos a un sistema de indexado, debido principalmente a la gran dimensión de los mismos[5-7]. La complejidad computacional para la indexación y recuperación de imágenes de rostros hacen necesario crear sistemas de recuperación eficientes y escalables.

La tarea de **recuperación de imágenes de rostros** está estrechamente vinculada a técnicas de **recuperación de imágenes basadas en contenido** (CBIR por sus siglas en inglés) a partir de rasgos como el color, la textura o la forma[8]. Aunque los sistemas de CBIR existentes obtienen buenos resultados en la recuperación de imágenes, el rendimiento de los mismos se ve afectado cuando es aplicado a imágenes de rostro[7, 9]. Esto se debe principalmente a que los rasgos utilizados no resultan muy discriminativos y no se tienen en cuenta otros que son fundamentales en la representación de un rostro[7, 9].

Existen en la literatura varios algoritmos que tienen como propósito dar solución a la recuperación de imágenes de rostros, algunos basados en la semántica (palabras claves o metadatos) y otros en el contenido de la imagen (color, textura, forma, etc.). Los sistemas basados en la semántica permiten reducir el espacio de búsqueda a partir de la descripción de los rasgos de un rostro, por ejemplo: color del pelo, forma y tamaño de la nariz, etc. Estos métodos pueden ser empleados como primer paso en un proceso de reconocimiento facial, y permiten mediante la reducción del espacio de búsqueda lograr una mayor eficiencia [10]. Sin embargo estos algoritmos necesariamente traen consigo subjetividad e imprecisión. La subjetividad se debe a los diferentes puntos de vista de los usuarios al describir un rostro. La imprecisión es causada por lo complejo que resulta la cuantificación y calificación de un rasgo facial, por ejemplo:(i) decidir cuándo calificar una nariz como “ancha” y (ii) cuantificar “cuan ancha” es la misma [11].

Entre los algoritmos basados en el contenido de la imagen, algunos enfocan la recuperación en la selección de rasgos discriminativos, la reducción de la dimensión de estos y el empleo de un clasificador que permita obtener los mejores candidatos, reduciendo el espacio de búsqueda [12]. Existen otros métodos que además de emplear rasgos discriminativos para la representación de la imagen y reducir los mismos con el objetivo de lograr eficiencia y escalabilidad, utilizan una estructura de indexación que permite la recuperación de los mejores candidatos de forma rápida y efectiva. Este trabajo se enfoca en estos últimos, los sistemas de **recuperación de imágenes de rostros basados en contenido** (RIRBC) mediante el empleo de **estructuras de indexación**.

En este reporte no se pretende realizar un estudio de las técnicas y estructuras de indexación existentes, sino hacer un análisis profundo de los métodos reportados en la actualidad para la recuperación de imágenes de rostros basados en contenido, sus fortalezas y debilidades, de manera que se puedan detectar posibles líneas abiertas de investigación en esta temática. No obstante, primeramente

se resumen en la Sección 2, las estructuras de indexación más utilizadas en esta área y luego, en la Sección 3, se detallan los principales métodos de RIRBC reportados en la literatura.

2 Estructuras de indexación utilizadas en la recuperación de imágenes de rostro

A medida que el tamaño de una base de datos de imágenes aumenta, la velocidad de la recuperación es un aspecto importante a tener en cuenta. En este sentido, métodos de indexación para datos de altas dimensiones como los rasgos comúnmente extraídos de las imágenes se hacen cada vez más necesarios, ya que en este dominio los algoritmos de indexación tradicionales dejan de ser efectivos por el problema conocido como "*the curse of dimensionality*", que implica que su rendimiento se degrada a medida que la dimensionalidad del espacio de características aumenta [13]. Así han surgido estructuras de indexación específicas para datos de esta naturaleza [14].

La mayoría de los sistemas de RIRBC reportados en la literatura, no se enfocan en diseñar una estructura de indexación específica para las imágenes de rostros, sino que emplean alguno de los esquemas existentes. Las estructuras de indexación más empleadas en los algoritmos que se describirán posteriormente en la Sección 3 son: ***vantage object***, ***kd-tree***, ***hash table*** e **índice invertido**. Para un estudio más profundo de las diferentes técnicas de indexación existentes puede consultarse [15].

Existe un gran número de investigaciones sobre el alto rendimiento del **índice invertido** [16] ya que es la estructura de datos más popular en sistemas de recuperación de documentos como motores de búsqueda. Es una estructura de datos indexada que almacena la dirección de un dato (palabra, número, etc.) en una base de datos, documento o grupo de documentos. La mayoría de los documentos se almacenan como listas de palabras, mientras que los índices invertidos almacenan para cada palabra la lista de los documentos en los que esta aparece. Existe una gran variedad en los índices invertidos, estos no solo almacenan los documentos en los que una palabra aparece, además pueden registrar la frecuencia, la posición en el documento, ya sea en la oración, la línea, el párrafo o página. Incluso se encuentra en el caso en el que se construyen más de un índice invertido, por ejemplo en uno se almacena la palabra y su frecuencia de aparición, mientras en otro se almacena la posición. Otra posible variación es si el ***lexicon*** se almacena de forma separada o no. El *lexicon* almacena todas las muestras (palabras individuales o distintivas) indexadas para toda la colección de documentos, usualmente también almacena información estadística de cada muestra, como el número de documentos en los que aparece. Debido a la diversidad en la implementación de los índices invertidos el espacio empleado por estos varía entre un 5 y un 100% del tamaño total de los documentos indexados.

as **tablas *hash*** [17-19] permiten uno de los tipos de búsquedas más eficientes: ***hashing***. Una tabla *hash* consiste en un arreglo en el que se accede a la información mediante un índice especial denominado llave (***key***). La idea principal es establecer un mapeo entre el conjunto de todas las posibles *keys* y las posiciones en el arreglo empleando una **función *hash***. Una función *hash* acepta una *key* y retorna un **código *hash*** o **valor *hash*.** El tipo de valor de las *keys* puede variar, pero los códigos *hash* siempre son enteros. Es común que el número de entradas en una tabla *hash* sea relativamente menor al universo de posibles *keys*. Debido a esto la mayoría de las funciones *hash* mapean algunas *keys* a la misma posición por lo que ocurre una **colisión.** Una función *hash* eficiente minimiza el número de colisiones, aunque es necesario estar preparado para lidiar con ellas. Las tablas *hash* se pueden aplicar en una gran cantidad de escenarios, como en sistemas de bases de datos, específicamente aquellos que requieren acceso aleatorio eficiente. Generalmente un sistema de base de datos trata de optimizar entre dos tipos de acceso: aleatorio y secuencial. Las tablas *hash* juegan un rol importante en el acceso aleatorio eficiente, ya que brindan una vía para localizar información en un espacio de tiempo constante. Otra aplicación de las tablas *hash* que merece la pena mencionar es su uso en la construcción de diccionarios de datos. La tabla *hash* es una estructura de datos que soporta la adición, substracción y

búsqueda de información. Además las operaciones de una tabla *hash* y un diccionario de datos son semejantes, lo que hace que su empleo sea particularmente eficiente.

Un árbol ***kd-tree*** [20, 21] es una estructura de datos que se emplea para almacenar una cantidad finita de elementos en un espacio de dimensión k. Este es un árbol binario que está diseñado para manipular de forma simple conjuntos de datos de gran dimensión. En un conjunto de datos E representados en un *kd-tree* cada elemento E_i es almacenado en un nodo. Cada nodo contiene dos punteros que pueden ser nulos o señalar otro nodo. Existen varios métodos para la construcción de un árbol *kd-tree*, aunque de forma básica se busca la media en los datos del conjunto E, la cual se selecciona como nodo raíz. Esto hace que por ejemplo en el caso de trabajar con puntos, todos los valores menores que el nodo raíz aparecerán en el subárbol izquierdo y los valores mayores en el izquierdo. No es necesario realizar la búsqueda de la media en el conjunto de datos para la construcción del árbol, aunque esto hace que el árbol no sea balanceado y afecte su eficiencia. Existen diferentes métodos para realizar búsquedas en el *kd-tree* como el **vecino más cercano** (*NN* por sus siglas en inglés) o la **búsqueda por rango** [21].

Vantage object [22] es un paradigma para almacenar información en una estructura de datos de forma tal que objetos con características similares puedan recuperarse de forma eficiente. Esta estructura se emplea en la recuperación de imágenes y se basa en una función de distancia para obtener la similitud entre dos imágenes. Para cada objeto (rasgo) de una imagen en una base de datos, se calcula su distancia a un conjunto predeterminado de m objetos con posición de superioridad (*vantage objects*); este vector de distancias representa un punto p en un espacio *vantage* de dimensión m. Los objetos de una base de datos que son similares (atendiendo a la función de distancia) a un objeto de consulta se pueden determinar mediante una búsqueda eficiente del vecino más cercano en el espacio *vantage*. Si se consideran dos objetos similares A_1 y A_2 (la distancia $d(A_1, A_2)$ es pequeña) y un tercer objeto A^* que se denomina *vantage object*; se puede medir el parecido entre los objetos A_1 y A_2 al comparar sus respectivas distancias a A^*, o sea si $|d(A_1, A^*) - d(A_2, A^*)|$ es pequeña A_1 y A_2 son parecidos. Dado un objeto de consulta $A_?$ se puede determinar el conjunto de objetos similares a este (i) calculando el vector de distancias $p_?$ entre él y los m *vantage objects* definidos, y (ii) seleccionando todos los objetos que poseen un vector de distancias $p_i = \{x_1, \dots, x_m\}$ similar a $p_?$. Si se formaliza lo antes expuesto cada objeto A_i es representado mediante un punto p_i en el espacio *vantage* definido por $A^* = \{A_1^*, \dots A_m^*\}$.

3 Algoritmos de recuperación de imágenes de rostros basados en contenido que utilizan estructuras de indexación

Los sistemas de RIRBC se pueden clasificar en dos categorías, teniendo en cuenta los rasgos de bajo nivel empleados para la representación de las imágenes de rostro: sistemas basados en la forma y sistemas basados en la apariencia. Los métodos de reconocimiento facial basados en la forma (características geométricas) fueron populares durante los inicios del reconocimiento automático de rostros, a finales de 1960 y principios de 1970 [23]. Estos métodos se basan en los puntos característicos (fiduciales) de un rostro, sus propiedades y las relaciones que existen entre ellos (distancias, ángulos, áreas) [4]. Son empleados frecuentemente para resolver problemas de pose e iluminación [24], aunque dependen de imágenes con buena calidad y de la correcta detección de los rasgos faciales. Estos métodos además desprecian gran parte de la información contenida en un rostro, como la textura o el color. Por otro lado, los métodos basados en la apariencia consideran directamente los valores de intensidad de los pixeles en la imagen y a partir de estos obtienen representaciones de la textura y rasgos faciales [4]. Desde 1990 estos métodos se han convertidos en los más empleados para el reconocimiento facial, debido a que han elevado la efectividad y la eficiencia en esta tarea [24].

Los métodos basados en la apariencia se pueden emplear de dos formas diferentes: holística (global) o local. Los métodos holísticos identifican un rostro a partir de un vector de rasgos que representa la imagen completa, mientras que los locales emplean vectores de rasgos que representan diferentes regiones de la imagen del rostro [4, 25]. Los rasgos locales en comparación a los globales son menos afectados por problemas de iluminación y pose, además son más flexibles en cuanto a la detección del rostro en la imagen [25] y han mostrado mayor eficacia en el reconocimiento [24].

La mayoría de los sistemas de RIRBC utilizan rasgos de apariencia local, extraídos de regiones específicas del rostro como los ojos, la nariz y la boca [2, 5, 7, 9]. Partiendo de estos rasgos se puede proponer el uso de una estructura de indexación directa (**indexación de los rasgos)** o se puede obtener una **representación de la imagen a alto nivel.** Las propuestas más recientes se basan en esta última estrategia y para lograr la representación de la imagen a alto nivel se utilizan dos vías fundamentales: los métodos de **bolsas de palabras** (conocidas como ***BoW*** por sus siglas en inglés) [26] y las **representaciones dispersas** (conocidas como ***sparse representation*** en inglés**)** [27]. Para una mayor comprensión de lo antes expuesto, se puede observar en la Fig. 1 una taxonomía propuesta para clasificar los métodos de RIRBC existentes que utilizan estructuras de indexación.

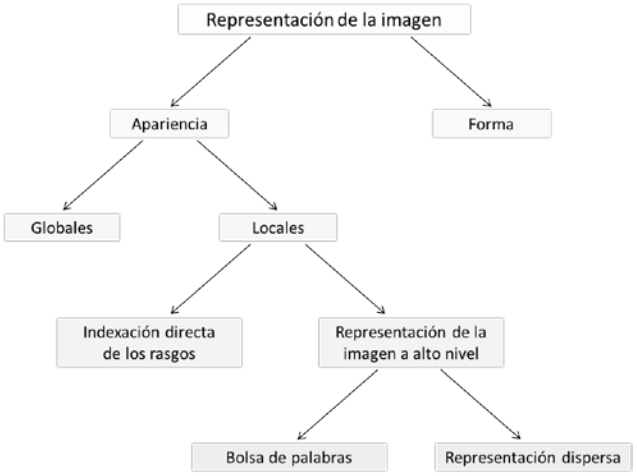

Fig. 1. Taxonomía de los métodos de RIRBC.

En el caso del empleo de representación de la imagen basado en la forma o en la apariencia mediante rasgos globales se procede directamente a realizar la indexación. La cual también es necesaria luego de representar la imagen a alto nivel mediante *BoW* o representación dispersa.

En los siguientes epígrafes se explican de forma detallada el empleo de estas estructuras de indexación y algunos de los métodos más representativos. Es necesario resaltar, que la línea más seguida en la RIRBC son los métodos basados en apariencia local, con los que se ha obtenido la mayor eficiencia y eficacia en esta tarea en los últimos años.

3.1 Algoritmos basado en forma

Los algoritmos basados en la forma del rostro, se basan en mediciones realizadas a partir de puntos característicos del rostro, que necesitan ser localizados con exactitud. Por este motivo, no han sido muy desarrollados en las investigaciones recientes. En la literatura aparece un método de RIRBC basado en la forma del rostro propuesto por Vikram en [28] emplea la similitud espacial para el indexado de las imágenes de rostros. En este se identifican 18 puntos característicos en el rostro, siguiendo la nomenclatura de la base de datos BioId [29], ver Fig. 2. Las coordenadas de los puntos son

manualmente identificadas y etiquetadas de forma única para obtener la **representación simbólica** de un rostro. Cada punto significativo en el rostro es tratado como un objeto relevante en la representación simbólica. Luego, la dispersión espacial de los puntos significativos de los rostros se conserva en la estructura indexada *kd-tree* [21] para lograr una recuperación eficiente. La metodología propuesta tiene como objetivo (i) ser invariante a transformaciones lineales, y (ii) ser robusta a variaciones de pose y expresión.

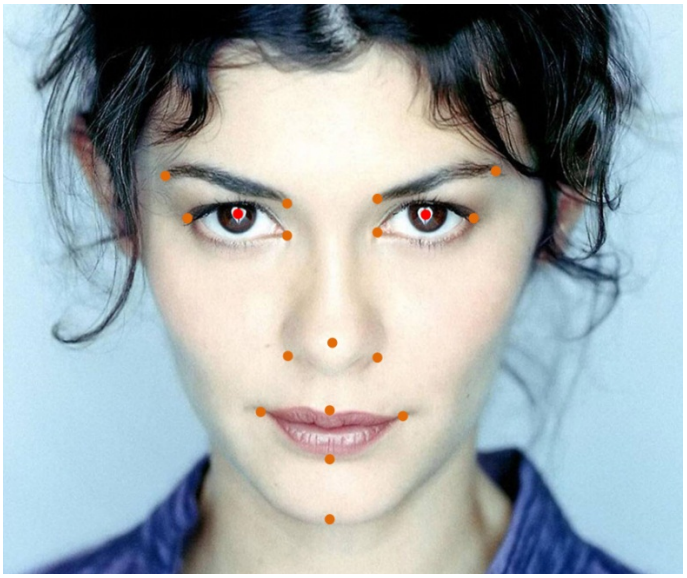

Fig. 2. Localización manual de 18 puntos significativos en el rostro.

Con el objetivo de preservar la dispersión espacial entre los objetos relevantes, se emplea la relación de distancia espacial normalizada. A partir de los objetos relevantes se construyen triángulos sobre el rostro, formando una especie de malla como se muestra en la Fig. 3. La relación espacial entre tres objetos relevantes A, B y C de una imagen se puede observar en la Fig. 4.

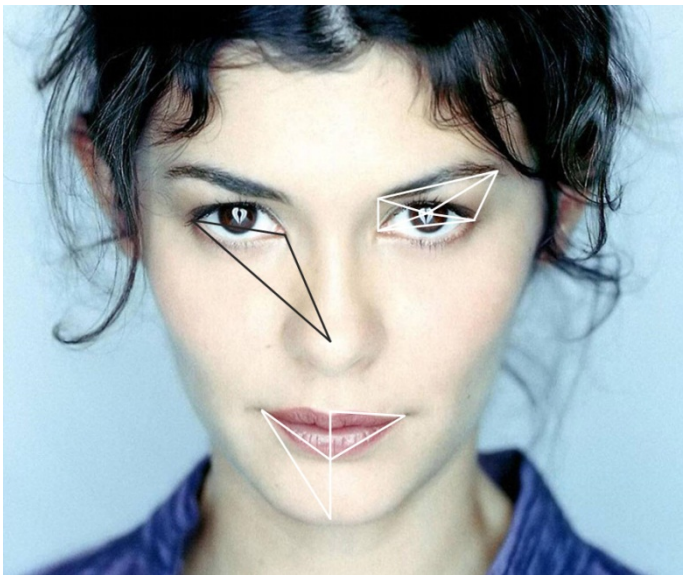

Fig. 3. Construcción de malla triangular sobre un rostro a partir de los objetos relevantes.

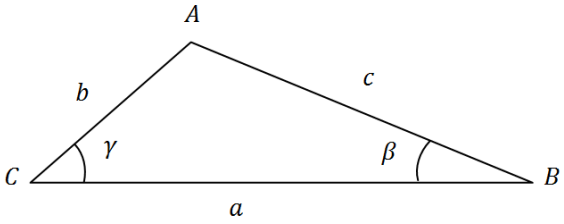

Fig. 4. Relación espacial entre los objetos relevantes ***A***, ***B*** y ***C***.

En la Fig. 4, a es el tamaño del lado mayor del ΔABC, b y c las distancias de los otros dos lados de triángulo. La dispersión intra-espacial se preserva al calcular los radios $\frac{b}{a}$ y $\frac{c}{a}$, referidos en este trabajo como "distancias normalizadas". Estas distancias preservan la dispersión intra-espacial de forma uniforme incluso cuando ΔABC es sometido a transformaciones lineales como rotación, traslación o escalado uniforme.

Cada triángulo es representado mediante una tupla de cinco elementos $(Lo_i, Lo_j, Lo_k, nd_j, nd_k)$ donde L_x representa la etiqueta cada objeto x: o_i, o_j y o_k. Los elementos nd_j y nd_k representan las distancias normalizadas de los lados opuestos a Lo_j y Lo_k. El orden de las etiquetas de los objetos en la tupla depende del tipo de triángulo formado por o_i,o_j y o_k. Si $(Lo_i, Lo_j, Lo_k, nd_j, nd_k)$ describen la tupla de un triángulo entonces tienen que satisfacer los siguientes criterios:

(i) $Lo_{j'} \geq Lo_{y'} \geq Lo_{z'}$, y $\Delta O_x O_y O_z$ es equilátero,
(ii) $Lo_{j'} \geq Lo_{y'}$, $Do_x = Do_y$, y $\Delta O_x O_y O_z$ es isósceles,
(iii) $Lo_{j'}, Lo_{y'}, Lo_{z'}$, tal que $Do_x > Do_y > Do_z$ y $\Delta O_x O_y O_z$ es escaleno,

donde Do_x, Do_y y Do_z representan la dimensión de los lados opuestos a los ángulos en O_i,O_j y O_k en $\Delta O_x O_y O_z$. La representación simbólica de una imagen no es más que el conjunto de tuplas obtenido a partir de los puntos significativos.

A partir de este proceso se construye una base de datos, con las representaciones simbólicas de las imágenes (conocida como **SFID** por sus siglas en inglés) y los respectivos índices de las mismas, que se almacena en un *kd-tree*. Por tanto la representación de la información en el *kd-tree* es $[Lo_i, Lo_j, Lo_k, nd_j, nd_k, índice\ de\ la\ imagen]$. La habilidad de *kd-tree* para facilitar la búsqueda por rango (*searching range*) y el excelente tiempo de recuperación, fueron los principales motivos en los que se fundamentó su elección como estructura de datos.

El proceso de recuperación se realiza mediante la obtención de las tuplas que representan a la imagen de consulta. Estas tuplas se transforman en rangos para facilitar la búsqueda, como se muestra en la Tabla 1.

Tabla 1. Representación de una tupla.

Límite inferior	L_{oi}	L_{oj}	$nd_j - \varepsilon_1$	$nd_k - \varepsilon_2$	l
Límite superior	L_{oi}	L_{oj}	$nd_j + \varepsilon_1$	$nd_k + \varepsilon_2$	u

Los valores ε_1 y ε_2 representan los umbrales con los que las distancias normalizadas son modificadas. Estos umbrales permiten que las distancias normalizadas varíen entre los valores 0 y 1, como se muestra en las restricciones 1 y 2. Los valores l y u denotan el rango del índice de la imagen en la base de datos SFID.

$$0 \leq nd_j - \varepsilon_1, nd_j + \varepsilon_1 \leq 1, \quad (1)$$

$$0 \leq nd_k - \varepsilon_2, nd_k + \varepsilon_2 \leq 1. \quad (2)$$

En la experimentación y validación realizada por Vikram se empleó la base de datos de rostros ORL [30]. La región del rostro en esta base de datos es recortada y redimensionada a $112\ X\ 92$ pixeles. Los 18 puntos característicos del rostro sugeridos son identificados manualmente tanto en la base de datos ORL original como en la modificada, y a continuación se crea la base de datos SFID. La base de datos ORL consiste en 400 imágenes de rostros de 40 sujetos, a razón de 10 imágenes por sujeto. Muestras alternas de cada clase son empleadas como conjunto de entrenamiento o de prueba.

Para comprobar la eficiencia del algoritmo de indexación propuesto, se evalúa el poder de indexación correcta (conocido como *C.I Power* por sus siglas en inglés) al recuperar la primera imagen relevante de un rostro a identificar, con un recorrido mínimo de la base de datos, ver Ecuación (3).

$$C.I\ Power = \frac{N_{ci}}{N_d}. \quad (3)$$

N_{ci}= Número de imágenes correctamente indexadas.
N_d = Número de imágenes en la base de datos.

El algoritmo de indexado propuesto se comparó con tres esquemas, propuestos por Gudivada [31], El-Kwae [32] y Sciascio [33], denominados como SIM_G, SIM_{DTC} y SIM_L respectivamente. En el primero Gudivada propone una estructura basada en la geometría para representar la relación espacial en imágenes y un algoritmo de similitud espacial. En el segundo El-Kwae introduce una firma denominada Two Signature Multi-Level Signature File (2SMLSF) como una estructura de acceso a grandes bases de datos de imágenes. En este algoritmo se codifica la información de la imagen en firmas binarias y se construye un árbol como estructura para realizar búsquedas de forma eficiente. En el tercer esquema, Sciascio propone un algoritmo basado en el cotejo de grafos. Como resultado de la comparación de estos esquemas se obtuvo que la precisión en la indexación es de un 90% para un 5% de la base de datos indexada recorrida, superando los otros esquemas.

Vikram además analiza el comportamiento del sistema atendiendo a los valores de *precision* y *recall*, mediante los cuales se puede comprobar la eficiencia del esquema de indexación propuesto. En este ámbito, *precision* (también llamado valor predictivo positivo) es la fracción de casos recuperados que son relevantes, mientras que *recall* (también conocida como la sensibilidad) es la fracción de casos pertinentes que se recuperan. A medida que aumenta el valor de *presicion* disminuye el *recall*, lo que converge en eficiencia al recuperar las imágenes relevantes. En la Fig. 5 (tomada de [28]) se puede observar que la metodología propuesta por Vikram mejora los esquemas de referencia.

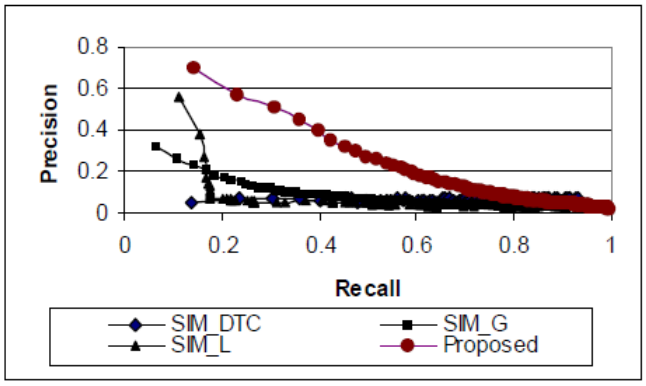

Fig. 5. Representación gráfica los parámetros *recall* y *precision* [28].

A pesar que el método propuesto por Vikram al usar la información geométrica contenida en el rostro ataca problemas comunes en el área del reconocimiento facial como: la variación en la pose y la expresión, y los cambios de iluminación; este no resulta ser del todo eficiente. Un inconveniente de este esquema es que los resultados están condicionados a la precisión en la localización de los 18 puntos en el rostro, proceso que en este algoritmo se realiza manualmente, lo que representa un contratiempo. Es cierto que existen algoritmos para la detección de estos puntos, pero están condicionados a la calidad de las imágenes para obtener buenos resultados. Sin embargo, el principal problema de esta propuesta es que no se tiene en cuenta gran parte de la información contenida en la imagen de un rostro como el color y la textura.

3.2 Algoritmos basados en apariencia global

Los métodos de enfoque holístico presentan dos ventajas fundamentales: (i) preservan la información de la textura y la forma y (ii) capturan más aspectos globales del rostro que los rasgos locales [24]. Sin embargo, existen pocos algoritmos de RIRBC propuestos para este tipo de rasgos, que son los menos utilizados en la actualidad en el Reconocimiento de Rostros. El primero, de dos algoritmos encontrados

en la literatura, consiste en una etapa de un sistema propuesto por Wang [34] para la *anotación automática de rostros*. El segundo algoritmo, propuesto por Lin [5], se basa en los rasgos globales obtenidos mediante el análisis de componentes principales (conocidos como *eigenfaces*), que indexa los rostros utilizando ***vantage object*** [22] como estructura de indexación.

La *anotación automática de rostros* es un servicio de gran relevancia para redes sociales y dispositivos móviles, este consiste en etiquetar con un nombre las imágenes de los rostros de manera automática [34]. Wang [34] propone un algoritmo para la anotación de rostros basado en la recuperación de imágenes (conocido como RBFA por sus siglas en ingles). En este método Wang propone una etapa para el pre-procesamiento de las imágenes e indexado de rasgos de gran dimensión. Para la representación de la región facial extraída se extraen rasgos globales GIST [35], que mediante el esquema de búsqueda aproximada *Locality-Sensitive Hashing* (LSH) [17] son indexados. En la etapa de RIRBC a partir de imagen de consulta se realiza un procedimiento similar para obtener, mediante la técnica de LSH, una pequeña lista de los rostros más similares de la base de datos. Según Kafai [14] los métodos LSH presentan como limitación común el requerimiento de que los puntos sean representados por un vector explícito. Además el empleo de los rasgos globales GIST, que fueron preconcebidos para la tarea de reconocimiento de escenarios mediante la clasificación de imágenes, trae consigo que no se tenga en cuenta información propia de un rostro. Estos rasgos globales representan un inconveniente debido a que se ven afectados por variaciones de pose e iluminación, son más sensibles a errores en la detección del rostro en la imagen y no son lo suficientemente discriminativos [4].

Existen varias técnicas de reconocimiento de rostros a partir de rasgos globales que emplean Análisis de Componentes Principales (PCA por sus siglas en inglés) como base [4]. Entre estas técnicas la representación de rostro más empleada es la denominada *eigenfaces*, desarrollada por Turk y Pentland para la detección e identificación de rostros [25]. Basado en esta representación de los rostros, Lin [6] propone un esquema para indexar rostros que utiliza ***vantage object*** [22] como estructura de indexación.

El método PCA tiene como objetivo extraer información relevante en la imagen de un rostro y representarla de forma eficiente. La idea es capturar las variaciones en una colección de imágenes de rostros y sobre esta base extraer la información contenida en cada imagen de forma individual. Esto consiste en encontrar los *eigenvectors* de la matriz de covarianza del conjunto de imágenes, o sea los componentes principales en la distribución de rostros. Los *eigenvectors* caracterizan las variaciones entre imágenes de rostros. Cada imagen contribuye en cierta medida a un *eigenvector*, por lo que un *eigenvector* se puede mostrar como un tipo de rostro fantasmal denominado *eigenface*, como se puede apreciar en la Fig. 6. Las imágenes del conjunto de entrenamiento pueden representarse como una combinación linear de *eigenfaces*. Esta técnica reduce dimensionalidad atendiendo al número de *eigenfaces* empleados para representar la imagen. La imagen de entrada se proyecta en el *eigenspace* para formar un vector de rasgos para su representación y posterior reconocimiento.

Fig. 6. Rostros fantasmales conocidos como *eigenfaces*.

En el sistema propuesto por Lin [6] se obtienen los *eigenfaces* a partir de una base de datos de imágenes de rostros. Cada rostro en la base de datos es ordenado a partir de su proyección en cada uno de los *eigenfaces*. Al identificar un rostro, la imagen de entrada es ordenada de forma similar, y los rostros más cercanos a su posición con respecto a los *eigenfaces* son seleccionados de la base de datos. Esta selección conforma una pequeña base datos, denominada *base de datos densa*, que se utiliza para el reconocimiento facial, en lugar de considerar la base de datos original.

Existen esquemas de indexación basados en estructuras de árboles que pueden emplearse para vectores de rasgos de tamaño menor que 20 [6]. Sin embargo, para la tarea de reconocimiento facial la dimensión de un vector de rasgos es mucho mayor, a pesar de que al utilizar PCA esta dimensionalidad se reduce. En el esquema propuesto por Lin, se propone entonces emplear los *eigenfaces* para formar una estructura ***vantage object,*** que permite seleccionar eficientemente rostros similares en una base de datos. De esta forma se puede (i) reducir el problema de lidiar con grandes volúmenes de imágenes y, (ii) en una segunda etapa, emplear un método más preciso para el reconocimiento facial.

En la estructura de indexación propuesta, la magnitud de cada proyección k_i, se usa de forma individual para ordenar las imágenes de los rostros en p listas ordenadas, siendo p el número de *eigenfaces*.

El costo requerido para buscar la imagen de un rostro en una base de datos, depende entonces de la dimensión del vector de rasgos y del número de sujetos en la base de datos. El proceso de búsqueda es más rápido si un sub-conjunto de la base de datos es seleccionado para confeccionar una *base de datos densa* que incluya el rostro buscado.

Para conformar la *base de datos densa*, cada imagen de la base de datos es normalizada, sustraída del rostro promedio y proyectada en los p *eigenfaces* con los *eigenvalues* correspondientes. Cada imagen entonces es compuesta de p valores proyectados $[k_{i,1}, k_{i,2}, \dots, k_{i,p}]$. A partir de cada *eigenface* se organizan los rostros de forma ascendente para formar P listas ordenadas. En la Fig. 7 se ilustra el esquema de indexación propuesto con p *eigenfaces*. El valor proyectado en el m^{th} *eigenface* con una posición j es denotado en el esquema como x_j^m. La posición j de un rostro en una lista ordenada p_i esta dada por la similitud del mismo respecto al *eigenface* m^{th}. Rostros semejantes tienen posiciones cercanas respecto a los *eigenfaces*.

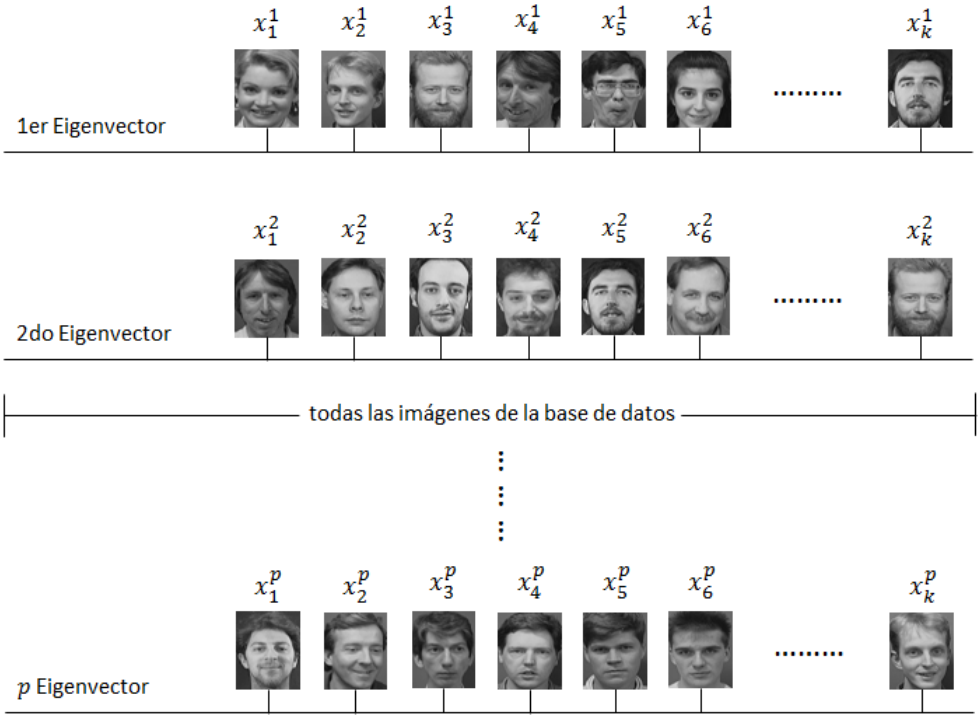

Fig. 7. Esquema de indexación propuesto.

En el método propuesto, la imagen de entrada es normalizada y posicionada respecto a cada *eigenface*. Si Y representa la imagen normalizada y sustraída, sus proyecciones en cada *eigenface* v_i se obtienen como se muestra en la Ecuación (4).

$$y_i = v_i^T \cdot Y, \text{donde } i \approx 1, \dots, p. \quad (4)$$

La entrada Y es posicionada en las p listas ordenadas, para luego a partir de los vecinos más cercanos en cada lista, seleccionar en la base de datos imágenes de rostros similares como se muestra en la Fig. 8. Si la imagen de entrada es posicionada entre la posición j y $j+1$ en la lista m^{th}, la imagen de la base de datos que su valor proyectado x_j^m o x_{j+i}^m sea el más cercano a y_m es seleccionado y agregado a la base de datos densa *B*. Se seleccionan rostros similares de la base de datos al analizar cada una de las p listas ordenadas, este proceso se repite hasta que *B* contenga el número de rostros distintivos requeridos.

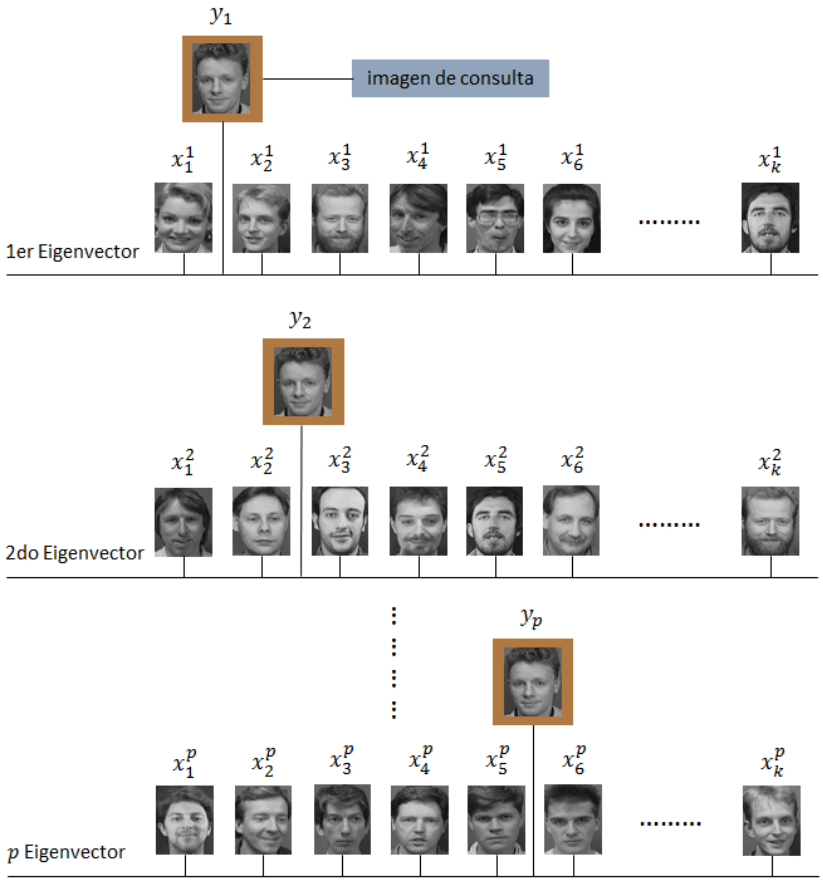

Fig. 8. Posicionamiento de una imagen de entrada, y selección de las imágenes para la construcción de la base de datos densa.

El costo computacional del esquema propuesto depende del número de *eigenfaces*. Un número pequeño de *eigenfaces* conlleva a un menor costo computacional, pero a su vez implica construir una base de datos densa *B* mayor, para evitar la exclusión del rostro buscado. El número de *eigenfaces* y el tamaño de la base de datos densa están dependen del tamaño de la base de datos de imágenes con la que se decida trabajar.

En la experimentación realizada por Lin, para evaluar el rendimiento del esquema de indexación, se emplearon imágenes de varias bases de datos y un grupo de imágenes capturadas por los autores de la investigación. Entre las bases de datos empleadas se encuentra: ORL[30], AR [36], BioID [29], UMIST [37] y de las universidades Bern [38], Yale [39] y MIT. A partir de las imágenes capturadas por los autores y las bases de datos, se conformó un set de 523 sujetos con un total de 752 imágenes de rostros para la experimentación. Las imágenes de cada sujeto presentan diferentes condiciones de iluminación, variaciones de pose o fueron tomadas en diferentes edades.

El principal objetivo de la experimentación fue, a partir del tamaño de una base datos de rostros, definir el número de *eigenfaces* y la dimensión de la base de datos densa para alcanzar el mejor rendimiento en términos de costo computacional y tasa de reconocimiento. Para ello se estudió el efecto de la cantidad de imágenes de la base de datos en el número óptimo de *eigenfaces* a usar y en el tamaño de la base de datos densa. Se emplearon dos tamaños de base de datos, 330 y 523.

En la experimentación se analiza el impacto que tienen diferentes números de *eigenfaces* sobre el tamaño de la base de datos densa, a partir de las diferentes bases de datos de rostros. Los resultados muestran como el tamaño de la base de datos densa disminuye a medida que aumenta el número de

eigefaces hasta cierto punto. Como resultado se obtuvo que el número óptimo de *eigenfaces* es aproximadamente el 25 % de las imágenes de rostros de la base de datos a emplear, mientras que el tamaño óptimo de la base de datos densa es un 35% de la base de datos original.

Lin se puede considerar como uno de los pioneros en la recuperación de imágenes de rostros a partir del empleo de estructuras de indexación. En el algoritmo propuesto, Lin demuestra que la reducción de la dimensionalidad del vector de rasgos que describe un rostro no es la única vía para reducir el espacio de búsqueda en el proceso de recuperación de imágenes faciales. Muestra además como con el empleo de los *eigenfaces* y la estructura de *vantage object* se puede reducir el espacio de búsqueda y lograr escalabilidad y eficiencia. Es importante reconocer que el algoritmo propuesto presenta problemas, entre ellos la elección de rasgos globales para describir un rostro y la falta de resultados experimentales. Los rasgos globales representan un inconveniente debido a que se ven afectados por variaciones de pose e iluminación, son más sensibles a errores en la detección del rostro en la imagen y no son lo suficientemente discriminativos [4]. Debido a esto se puede pensar que el empleo de rasgos locales podría converger en la construcción de una base de datos densa más representativa, y por ende evitar falsos positivos en el conjunto de rostros candidatos. Una experimentación un poco más consciente que incluya una comparación con un algoritmo de reconocimiento facial que realice una búsqueda lineal ayudaría a evidenciar la importancia de la metodología propuesta por Lin.

3.3 Algoritmos basados en apariencia local

La línea más seguida en la RIRBC son los sistemas basados en apariencia local. Estos sistemas se dividen en dos grupos, los que proponen una indexación directa de los rasgos y los que obtienen una representación de la imagen a alto nivel a partir de estos. El último enfoque es el más reciente y con el que mejores resultados se han obtenido. Partiendo de este enfoque se han desarrollado varias propuestas de algoritmos que emplean el índice invertido como estructura de indexación.

3.3.1 Indexación directa

Kaushik presenta en [40] un **esquema para el indexado** de una base de datos de rostros mediante una modificación a la técnica de ***geometric hashing***. En este sistema se extraen **puntos de control** (*pc*) que representan rasgos faciales de un rostro mediante el operador conocido como SURF, por sus siglas en ingles [41]. Además, propone un pre procesamiento de estos *pc* para conseguir que sean invariantes a la traslación, rotación y escalado. El método *geometric hashing* modificado se emplea para generar a partir de las coordenadas de los *pc* una tabla ***hash***.

En una imagen biométrica existen puntos de control (*pc*) que brindan características únicas. La información contenida en estos *pc* puede ser representada mediante una *s*-tupla. En el diseño de un esquema de indexación eficiente para una base de datos biométrica hay que tener en cuenta que (i) el número de *pc* en los rasgos biométricos es elevado y puede ser variable, (ii) la imagen a identificar puede estar rotada y trasladada con respecto a las de la base de datos, y (iii) la calidad de las imágenes puede ser pobre, por lo que los vectores de *pc* pueden contener falsos puntos o no incluir positivos.

La mayoría de los esquemas de indexación existentes trabajan a partir tamaños fijos de los vectores [40]. Sin embargo, en el caso de rasgos biométricos la cantidad de *pc* varía entre una imagen y otra. En este trabajo dan solución a este problema mediante el uso del método *geometric hashing*. Este método es un enfoque de reconocimiento de objetos basado en un modelo, en el que el objeto es representado mediante *pc* discretos como se muestra en la Fig. 9. El reconocimiento se realiza a través del macheo de los *pc* del objeto de entrada (modelo de entrada "*men*") con los predefinidos en la base de datos (modelo de base de datos "*mbd*"). Un *men* o un *mbd* de un objeto no es más que un grupo de *pc* llamados "puntos de interés" junto con su posición geométrica.

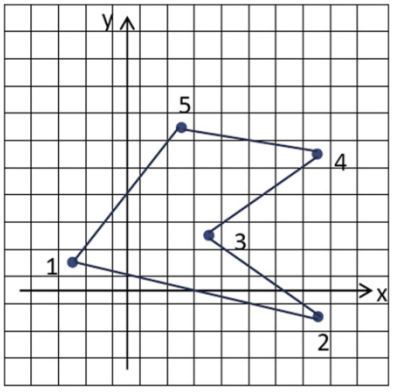

Fig. 9. Representación de un objeto de 5 puntos [40].

En ese trabajo se utilizó SURF para detectar los *pc* junto con sus descriptores, el cual ha mostrado buenos resultados en la descripción de imágenes de rostro [42]. Con este método, cada imagen puede representarse como un conjunto F de m puntos de control, $\{f_1, f_2 \ldots f_m\}$ y cada punto de control f_l es representado como una tupla de tres elementos (x_l, y_l, D_l) donde (x_l, y_l) son sus coordenadas en los ejes X, Y, y D_l es el descriptor correspondiente. Las coordenadas se emplean para la construcción de la tabla *hash* durante la indexación y los descriptores se utilizan en el reconocimiento. Con el objetivo de hacer los *pc* de cada imagen invariantes a rotaciones y escalas se realiza un pre procesamiento que consiste en los pasos relevantes (i) centrado medio de los *pc* (cada *pc* es trasladado de forma tal que la media de todos es 0), (ii) rotación de los *pc* mediante el uso de PCA, y (iii) normalización para hacer al método *geometric hashing* invariante al escalado.

Luego, en la fase de reconocimiento se realiza una búsqueda en la tabla *hash* generada contra los *pc* entrados para una nueva imagen. La tabla *hash* está compuesta por *bins* que contienen el identificador de un rostro junto con el *pc* y su descriptor. Este proceso consta de dos etapas, en la primera los *mbd* diferentes al *men* son descartados, mientras que en la segunda etapa se realiza una votación con los *mbd* seleccionados y se obtiene los *k* mejores resultados. En la primera etapa la similitud de un *pc* se decide a partir de la coordenada de posición y el *bin* en la tabla hash al cual es mapeado. Con el objetivo de mejorar el rendimiento del algoritmo la búsqueda se realiza no solo atendiendo al *bin* al que el *pc* fue mapeado, sino además a una vecindad de *bins* de tamaño K.

Para determinar el rendimiento del esquema de indexación propuesto, Kaushik empleó cuatro medidas: ***Hit Rate***, ***Bin Miss Rate***, ***Penetration Rate*** y ***Cumulative Match Characteristic curve***. *Hit Rate* es la taza de imágenes identificadas correctamente entre las *k* mejores coincidencias. *Bin Miss Rate* es $100\% - \text{Hit Rate}$, ambas mediciones dependen del tamaño de la vecindad de *bins* a analizar, mientras mayor la vecindad mayor *Hit Rate* y menor *Bin Miss Rate*. *PenetrationRate* es el espacio promedio de la base de datos que fue necesario emplear para realizar la búsqueda de una imagen a identificar. *Cumulative Match Characteristic curve* (CMC) es la relación entre el *Hit Rate* y el *Rank* (mejores k coincidencias).

En los experimentos Kaushik emplea la base de datos FERET [43], precisamente 206 imágenes de rostros de 103 sujetos con variaciones en la pose e iluminación. De cada sujeto se emplea una imagen para conformar la base de datos, mientras que las demás se utilizan para la tarea de identificación. Durante el desarrollo de los experimentos el autor asume que el tamaño de cada *bin* es infinito.

Entre los experimentos realizados se evaluó el *Hit Rate* con diferentes umbrales, para las *k* mejores coincidencias con diversas vecindades. El umbral definido en este experimento es la relación entre vecindad de *bins*, *Hit Rate* y las k mejores coincidencias. Los mejores resultados se obtuvieron con 4 mejores coincidencias y un umbral de 0.07, con un 100% de probabilidad.

La selección del tamaño de la vecindad de *bins* es crítica para obtener el mejor rendimiento de la técnica de indexación propuesta. Por lo tanto los experimentos fueron realizados a partir de varias vecindades, donde los correspondientes *Bin Miss Rate* y *Penetration Rate* obtenidos muestran que para esta base de datos el valor óptimo de vecindad es 4, con el cual se obtiene un *Hit Rate* de 100% para las 4 mejores coincidencias.

Kaushik comparó su algoritmo con dos esquemas de indexación convencionales basados en el método *geometric hashing*: *all posible triangulation* (APT) y la triangulación de Delauny (DT). Las comparaciones se realizaron a partir de la misma base de datos. Dos puntos similares a los *pc* de SURF y el vector descriptor a su alrededor de tamaño 64 se emplean en todos los esquemas de indexación comparados. Los resultados obtenidos muestran que el esquema de indexación propuesto por el autor presenta mejor rendimiento. Además se analizaron los valores de CMC para los tres esquemas, de los cuales el propuesto supera con creces a los demás.

La distribución del índice en la tabla hash es otro factor que se considera para comparar; el rendimiento del sistema de reconocimiento depende en gran parte de la distribución de los *pc*. Si la distribución de los *pc* es densa el tiempo de búsqueda es mayor, lo que afecta el rendimiento del sistema. Los resultados obtenidos por Kaushik muestran que en el esquema que propone los *pc* se encuentran bien distribuidos, casi todos los *bins* ocupan igual número de puntos. Razón por la cual, junto al hecho de que la cantidad de puntos en la tabla hash es menor que los otros esquemas, el autor define su algoritmo como eficiente.

El algoritmo propuesto por Kaushik tiene gran relevancia ya que en su metodología se tiene en cuenta gran parte de la información que se puede extraer de la imagen de un rostro para su representación. En este método no solo se emplea la información contenida en la imagen a nivel de pixel, extraída mediante el uso del descriptor SURF, sino que además a partir del empleo de la técnica de *geometric hashing* para la indexación, se tiene en cuenta información geométrica sobre las coordenadas de los puntos de control. A pesar que el descriptor SURF permite extraer rasgos locales de apariencia, es invariante a variación de escala y rotación, permite obtener información geométrica de la imagen y según Geng Du en [42] es adecuado para el reconocimiento facial, presenta algunos elementos que deben ser tomados en consideración. Entre las principales limitaciones resalta que no tiene en cuenta características faciales como los ojos, boca y nariz; simplemente localiza puntos de interés en la imagen y extrae los descriptores que describen a los mismos, desechando información importante para describir un rostro. Además en el estado del arte SURF se emplea de forma general para describir objetos al igual que el descriptor SIFT, al cual solo supera en velocidad de procesamiento [42]. También hay que resaltar que en el área del reconocimiento facial no se han realizado estudios que comparen su eficacia con otros descriptores que han tenido buenos resultados como LBP o HOG. Este esquema de indexación directa puede ser explorado con el uso de otros descriptores locales para determinar los puntos de control.

3.3.2 Representación de la imagen a alto nivel

En los últimos tres años se ha evidenciado un auge en la investigación y desarrollo de esquemas para la indexación y recuperación de imágenes de rostros, motivado por el uso de rasgos de alto nivel para representar el rostro. A partir de un primer estudio desarrollado por Wu [7] a finales del año 2011 se derivaron dos líneas fundamentales de investigación que plantean una representación de las imágenes de rostro a un alto nivel para su recuperación. La primera propone representar la imagen en palabras visuales mediante el método de **bolsas de palabras (*BoW*)**. La segunda línea de investigación adopta el método de **representación dispersa** (***sparse***) para la representación de la imagen, con el objetivo de dar solución a algunas restricciones presentes en el método *BoW*.

Estas líneas de investigación tienen en común la metodología para la extracción de rasgos locales establecida por Wu [7]. En esta metodología, el primer paso es la localización de cinco componentes faciales (dos ojos, la nariz y dos esquinas de la boca) y la normalización geométrica de la imagen. Luego se define una rejilla en cada componente detectado para obtener un total 175 parches de cada

rostro. De cada parche de la rejilla se obtiene un descriptor, contándose con un total de 175 descriptores por cada rostro. Además en todos estos trabajos [2, 5, 7, 9] se emplea como estructura de indexación el **índice invertido** [16].

Recuperación de imágenes de rostro usando bolsas de palabras BoW

Algunos de los sistemas de recuperación de imágenes que existen en el estado del arte alcanzan escalabilidad mediante la representación *BoW* y métodos de recuperación semántica. El rendimiento de estos sistemas se ve afectado en el dominio de imágenes de rostros, principalmente debido a que las palabras visuales generadas no tienen poder discriminativo para el caso de imágenes de rostros y se ignoran propiedades especiales de los mismos. Los principales rasgos empleados en el estado del arte para el reconocimiento facial pueden alcanzar buenos resultados, pero en general no son adecuados para el indexado inverso, ya que son de alta dimensión y globales, por lo que no son escalables en cuanto al costo computacional o de memoria [7]. En esta línea de investigación se han desarrollado dos trabajos fundamentales: el primero, percusor de una nueva tendencia en los esquemas de indexación de imágenes de rostros, desarrollado por Wu [7]. El segundo realizado por Zhen [9] en el año 2013, que tiene como objetivo mejorar el algoritmo propuesto por Wu.

Wu [7] propone un sistema de recuperación de imágenes de rostros escalable, partiendo de una nueva representación de rostro a partir de rasgos globales y locales. Este sistema consta de dos etapas principales: (i) indexación y (i) recuperación. En la primera etapa se explotan propiedades especiales de los rostros para extraer nuevos rasgos locales basados en componentes. Estos rasgos son cuantificados en palabras visuales mediante un nuevo esquema de cuantificación basado en identidad. Los rasgos globales discriminativos de cada rostro, por otra parte, son codificados a partir de una firma Hamming de 40 bytes. En la etapa de recuperación primero se obtienen imágenes candidatas del índice invertido de palabras visuales. Luego se utiliza una nueva distancia multireferencial para realizar un ranking de las imágenes candidatas mediante la firma de Hamming. En una base de datos de un millón de rostros se comprobó que los rasgos locales y las firmas globales de Hamming se complementan, ya que del índice invertido basado en rasgos locales se obtienen imágenes candidatas con buen *recall*, mientras que el ranking multireferencial con la firma Hamming global converge en un buen valor de *precision.*

En ese trabajo se asume que las imágenes de los rostros son frontales, con 20 grados de variación de la pose a lo sumo, de forma tal que los cinco componentes del rostro (ojos, nariz y esquinas de la boca) son visibles. El primer paso en la metodología propuesta es la extracción de rasgos faciales basados en componentes que sean robustos a variaciones de pose y expresión. Luego se realiza una cuantificación, en palabras visuales discriminativas, lo que permite indexar las imágenes de rostros, un paso crítico para lograr escalabilidad. Este proceso de cuantificación es basado en identidad ya que se le asigna de forma supervisada a cada parche extraído de la imagen un identificador. Por último además de los rasgos locales, se genera una firma Hamming de 40 bytes para cada rostro, lo que permite representar de forma compacta un rasgo global discriminativo de gran dimensión.

En la Fig. 10 se muestra la extracción e indexado de los rasgos faciales locales. Primero son localizados los principales componentes en el rostro mediante un detector de componentes basado en una red neuronal. Luego la imagen es geométricamente normalizada mediante el método trasformada de similitud que mapea la posición de los ojos a una posición canónica. Se define una rejilla de 5 X 7 regiones cuadradas en cada componente detectado, para obtener un total 175 parches en cada rostro; esto permite una deformación flexible entre los componentes y una mayor robustez ante variaciones de pose y expresión.

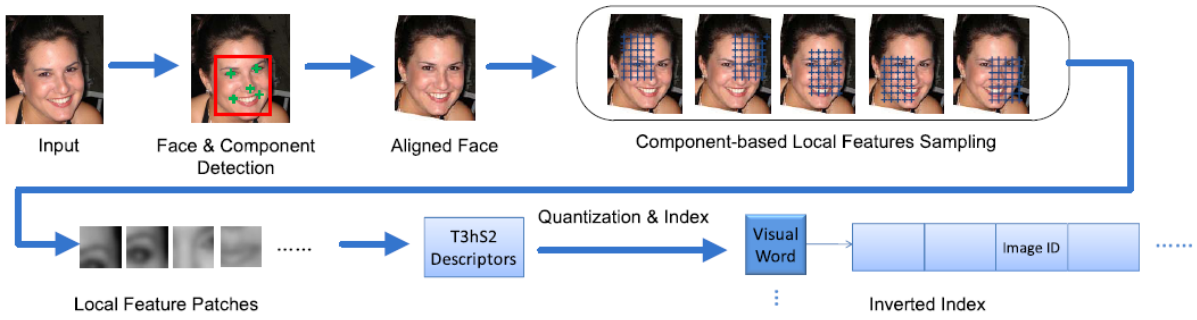

Fig. 10. Extracción e indexación de rasgos locales [7].

A partir de cada una de las regiones cuadradas de la rejilla se obtiene un descriptor T3hS2 [20], que es un descriptor local basado en filtros direccionales[44] y el descriptor de Histograma de Gradientes de Ubicación y Orientación (GLOH por sus siglas en inglés) [13]. Todos los descriptores son cuantificados en palabras visuales que son subsecuentemente indexadas. El uso de un descriptor basado en histogramas como T3hS2 y el hecho de que las rejillas de algunos componentes se solapan permiten al sistema tolerar algunos errores a la hora de localizar los componentes.

Con el objetivo de imponer restricciones geométricas entre los rasgos, se asigna manualmente a cada región de una rejilla un ID único llamado "id de posición". El id de posición se concatena con el id del rasgo de cuantificación para formar una "palabra visual". Dos rasgos pueden cotejar solo si provienen del mismo componente y son extraídos de la misma posición de la rejilla de ese componente.

Para lograr escalabilidad, los rasgos locales extraídos necesitan ser cuantificados en un grupo de palabras visuales empleando un vocabulario visual, que a menudo se obtiene mediante un algoritmo de agrupamiento como *k-means*. El aprendizaje no supervisado no es aconsejable para entrenar un vocabulario para el reconocimiento de rostros, donde las variaciones entre elementos de una misma clase son mayores que entre clases cuando los rostros presentan cambios de pose y expresión.

Wu presenta un modelo de cuantificación basado en identidad empleando aprendizaje supervisado. Los datos de entrenamiento consisten en P diferentes identidades con T imágenes de rostros con diferentes poses, expresiones y condiciones de iluminación. Debido a que cada persona tiene un "id de persona" y cada punto en una rejilla tiene un único "id de posición" se define la palabra visual como el par <id de persona, id de posición> asociado con los T descriptores de los rasgos locales generados de las muestras de entrenamiento de un "id de persona". La fortaleza de este esquema es que los rasgos bajo diferentes condiciones de pose, expresión e iluminación de una persona pueden ser cuantificados en una misma palabra visual como se muestra en la Fig. 11.

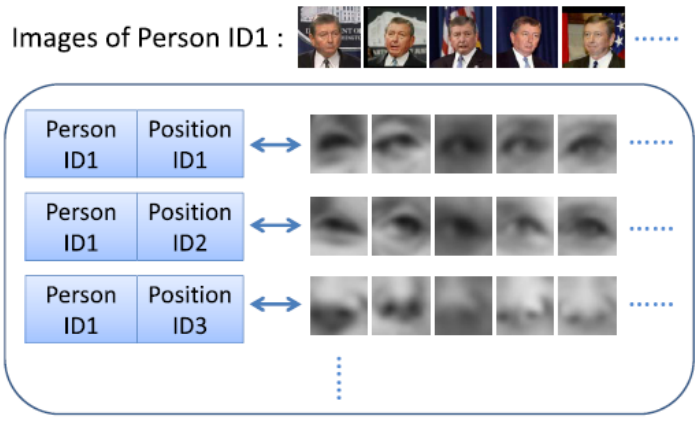

Fig. 11. Vocabulario de una persona basado en la identidad. Las palabras visuales se forman por <id de persona, id de posición> [7].

Con el vocabulario visual basado en identidad, la cuantificación es fácil de realizar mediante una búsqueda del **vecino más cercano** empleando árboles *k-d*. Por cada id de posición se crea un árbol *k-d* con todos los rasgos del conjunto de entrenamiento (T X P) como se muestra en la Fig. 12. Dado un rostro nuevo se localizan 175 parches, se extraen descriptores de cada uno y se hallan sus vecinos más cercanos de forma independiente.

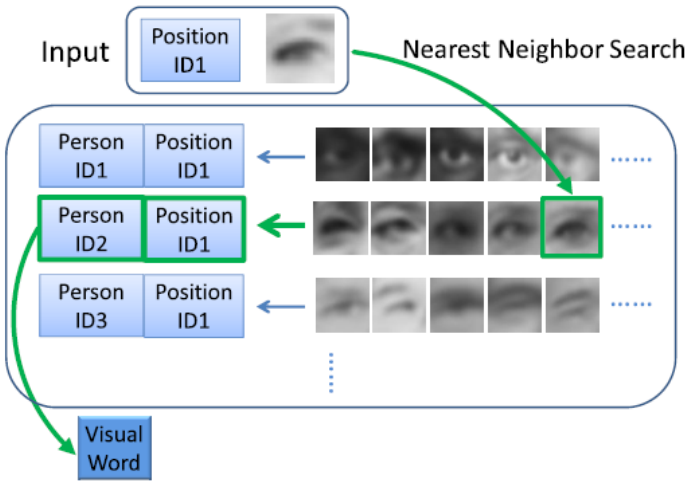

Fig. 12. Cuantificación de un rasgo extraído en la posición ID1 mediante una búsqueda del *vecino más cercano* [7].

El número de personas P y el número de muestras T influyen en la efectividad del vocabulario de tamaño (P X 175) y complejizan el sistema. Incrementar P es equivalente a aumentar el tamaño del vocabulario y por tanto el poder discriminativo de las palabras visuales. Sin embargo, aumentar T puede conducir a una mejor representación de un rostro, lo que puede reducir el número de errores de cuantificación. Es importante recalcar que al aumentar P y/o T se incrementa el costo computacional para realizar la cuantificación. En la experimentación los autores escogieron los valores de P y T empíricamente y se encontró que con $P = 270$ y $T = 60$ hay un mejor desempeño dado un consumo de memoria definido.

La idea básica del método propuesto para realizar el *reranking* es organizar los mejores candidatos mediante una pequeña firma extraída de rasgos globales. Esta signatura se obtiene mediante el uso de la firma Hamming, el descriptor *Learning-based* (LE) *Descriptor* [45] y PCA. A partir de esta firma se mejora la precisión sin comprometer la escalabilidad del sistema.

La firma de Hamming está basada en el descriptor LE, que se emplea junto con un codificador de proyección aleatoria basado en árboles, permitiendo convertir una imagen en un "código" de imagen. Cada imagen codificada se inserta en una rejilla de 5 x 7 celdas y de cada una de estas celdas se obtiene un histograma de 256 dimensiones, para al final obtener un histograma concatenado de 8,960 dimensiones. Este histograma final es comprimido mediante PCA para obtener como resultado un descriptor LE de 400 dimensiones.

Para crear la firma Hamming primero se crean aleatoriamente *Np* direcciones de proyección en el espacio del descriptor LE. Los descriptores LE de un conjunto de imágenes de entrenamiento son proyectados en cada una de las direcciones. La media de los valores proyectados representa el umbral para esa dirección. Mientras más proyecciones se empleen es posible una mayor aproximación. En este trabajo emplearon *Np* = 320 para obtener como resultado una firma de Hamming de 40 bytes, menor que el descriptor global LE en cuanto al almacenamiento y costo computacional.

A las imágenes candidatas obtenidas al recorrer el índice, se les realiza un *ranking* inicial basado en la cantidad de palabras visuales que cotejan con la imagen de la persona a identificar. Al lidiar con variaciones (pose, expresión o iluminación) entre muestras de una clase se emplea un grupo de imágenes de referencia para realizar un *re-ranking* a las imágenes candidatas. Se categoriza cada una de las imágenes candidatas a partir de su distancia promedio respecto a las imágenes de referencia.

Es importante la correcta selección de las imágenes de referencia, ya que podrían dañar el sistema. En este trabajo se emplea un enfoque iterativo para seleccionar las imágenes de referencia del grupo de imágenes candidatas. En cada iteración se selecciona una imagen que se parezca tanto a la de entrada como a las imágenes de referencias de la iteración anterior. Este enfoque permite un grupo de imágenes de referencias parecidas no solo a la imagen a identificar, sino además parecidas entre sí. Específicamente en cada iteración se selecciona una imagen *I* que minimice el siguiente costo:

$$D = d(Q, I) + \alpha \cdot \frac{1}{|R|} \sum_i d(R, I), \tag{4}$$

donde Q es la imagen de entrada, $\boldsymbol{R} = \{R_i\}$ es el set de referencia $d(\cdot,\cdot)$ es la distancia Hamming entre dos rostros, y α es un valor de peso. I se añade a $\boldsymbol{R}$ en un proceso iterativo que culmina cuando se seleccionan el número de imágenes de referencia deseado, o la distancia D es mayor que un umbral.

En la experimentación Wu empleó un detector de rostros para obtener un millón de rostros de la web, que sirven como base de datos de experimentación. Además añadió 40 grupos de imágenes de rostros de la base de datos *Labeled Face in Wild* (LFW) [46] (en total 1,142 imágenes) a la base de datos de experimentación para servir como *ground-truth*. Con el objetivo de evaluar la escalabilidad y el rendimiento en la recuperación creó tres bases de datos a partir de la utilizada para la experimentación, con imágenes de 10,50 y $200Kb$.

En la evaluación del rendimiento del sistema los autores seleccionaron 220 imágenes representativas del conjunto *ground-truth* para identificar. Como métrica para el rendimiento del sistema de recuperación se empleó la **media del promedio de precisión** (conocido como mAP por sus siglas en inglés), siendo en este caso el valor medio de los promedios de precisión de las 220 imágenes a identificar.

Como **línea base** Wu utilizó un enfoque en el que los rasgos locales se extraen de rejillas de $16\ x\ 11$ ubicadas sobre el rostro. El vocabulario visual se obtiene al aplicar el agrupamiento jerárquico k-*means* a 1.5 millones de descriptores de rasgos. La línea base es evaluada con dos tamaños de vocabulario: palabras visuales de 10 y $100KB$ para cada punto de una rejilla.

Los autores analizaron diferentes enfoques en la extracción de rasgos locales con varios métodos de cuantificación, entre los que el método basado en identidad propuesto mostró los mejores resultados. Se evaluó el *re-ranking* multireferencial y en los resultados se demostró que el sistema, con el empleo de la firma de Hamming y 10 imágenes de referencia (Nr), supera el enfoque de la línea base, con menor costo computacional y de memoria.

Otro de los aspectos analizados fue el impacto de la firma Hamming, que al aumentar su tamaño brinda mejor precisión en el *re-ranking*, aunque con mayor costo computacional. Entre los resultados obtenidos destaca que con un tamaño adecuado de código, el enfoque de *re-ranking* basado en la firma Hamming es significativamente más rápido que el enfoque basado en rasgos globales, con un rendimiento similar.

Para el análisis del impacto de los parámetros *Nr* y α en la selección de las imágenes de referencia para el *re-ranking*, Wu realizó una selección empírica empleando la base de datos de un millón de imágenes de rostros detectados en la web. Los resultados óptimos fueron $Nr = 10$ y $\alpha = 6.0$.

xisten dos parámetros que afectan de forma general el rendimiento del sistema: el rango de selección S y el rango de *re-ranking* M. El algoritmo propuesto para obtener el ranking selecciona las imágenes de referencias a partir de los S-mejores candidatos y realiza el *re-ranking* para obtener M-mejores candidatos. El parámetro S debería ser elevado, con el objetivo de cubrir la ocurrencia de candidatos positivos, aunque esta decisión trae consigo la inclusión de una mayor cantidad de falsos positivos al set de imágenes de referencia, lo que a su vez disminuye la precisión en la recuperación de los candidatos. Sin embargo la selección del parámetro M tiene como objetivo lograr un equilibrio entre el *recall* y el costo computacional. De los experimentos realizados por Wu se definieron como valores óptimos $S = 1000$ y $\mathrm{M} = 1000$.

Para evaluar la escalabilidad se analizó el costo computacional y de almacenamiento respecto al número de imágenes en la base de datos. El costo computacional de una búsqueda lineal se puede representar como $N \times D$, donde N es el número de imágenes y D la dimensión del rasgo global. En el enfoque propuesto, por cada rasgo local en la imagen a identificar, solo es necesario recorrer una pequeña parte del índice de palabras visuales. Se denota C al porciento del índice que se recorre, el cual que está relacionado con el tamaño del vocabulario. El número de rasgos locales extraídos de cada rostro se representa con N_F. El costo computacional del enfoque propuesto entonces se calcula como $C \times N_F \times N$ operaciones de voto. El valor de $C \times N_F$ es de uno o dos órdenes de magnitud menor que D, además la operación de votación es más rápida que el cálculo de la norma L1. En resumen el indexado de imágenes presenta una escalabilidad significativamente mejor que la búsqueda lineal en términos de costo computacional, ver Fig. 13.

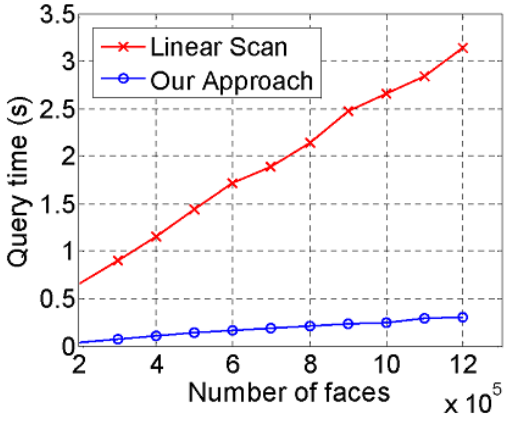

Fig. 13. Costo computacional [7].

Los autores además realizaron experimentos para comparar los resultados en la identificación de personas a partir de varios enfoques. Se comparó el método línea base de la cuantificación, la cuantificación basada en identidad, la búsqueda lineal con rasgos globales y el enfoque propuesto, la cuantificación basada en identidad con 10 imágenes de referencia para el *re-ranking* empleando firmas Hamming. En los resultados obtenidos se aprecia que el sistema propuesto es el mejor. En la Fig. 14 se puede observar un ejemplo de recuperación con los diferentes enfoques comparados.

Fig. 14. Ejemplo de los resultados de una búsqueda utilizando diferentes enfoques. A la izquierda se muestra el rostro que a buscar y a la derecha los mejores candidatos obtenidos de la recuperación. (a) Línea base de cuantificación. (b) Cuantificación basada en identidad. (c) Búsqueda lineal mediante rasgos globales. (d) Cuantificación basada en identidad junto con proceso de *reranking* empleando firmas Hamming [7].

Entre los aspectos más relevantes del algoritmo propuesto por Wu resalta la representación de la imagen de un rostro a partir del uso de rasgos globales y locales, esto permite analizar el rostro de una persona no solo a partir de componentes faciales específicos como la nariz, la boca o los ojos; sino además de forma holística. Es importante mencionar además que aunque el método de cuantificación basado en identidad resuelve el problema de la variación intraclase y permite lograr escalabilidad, este cuenta con el inconveniente de ser supervisado. La propuesta del método de *reranking* a partir de la construcción de una firma Hamming mediante rasgos globales es un aspecto muy importante en este algoritmo, ya que permite mejorar la precisión en la recuperación sin afectar la escalabilidad. Sin embargo el *reranking* depende de contar con una gran cantidad de imágenes de rostros por individuo para ser confiable, y además este método tiene un alto consumo de tiempo debido a que el proceso de

voto de rasgos locales es lineal [14]. También se puede mencionar que en lugar de emplear el descriptor T3hS2 para la construcción de las palabras visuales, se podrían utilizar LBP [4] o HOG [47], descriptores que han obtenido buenos resultados ante problemas de pose e iluminación. Este algoritmo tiene un gran impacto en el estado de arte ya que propone la metodología a seguir en las investigaciones que le suceden. Las bases de esta metodología son: (i) el empleo de rasgos globales y locales para describir la imagen de un rostro, (ii) la construcción de un diccionario de palabras visuales para indexar las imágenes y (iii) un proceso de *reranking* de las imágenes candidatas.

En un estudio más reciente Zhen plantea que una de las principales desventajas del algoritmo propuesto por Wu [7] es que no tiene en consideración la pose del rostro y su alcance es restringido a rostros frontales con una variación de hasta 20° de rotación. Debido a esto Zhen propone en [9] un algoritmo en el que se extraen rasgos locales basados en la pose, que son cuantificados en palabras visuales y empleados para la construcción de un índice invertido. Además se modifica el descriptor utilizado para el *re-ranking* con el objetivo de mejorar la eficacia en el proceso de recuperación mediante rasgos globales.

En el trabajo realizado por Wu se extrae un descriptor local de cada región cuadrada de cada rejilla. El rendimiento de este enfoque se ve afectado si el rostro no es frontal, esto se debe principalmente a que cuando el rostro se encuentra de perfil parte de alguna rejilla puede ubicarse fuera de la región del rostro. Para enfrentar este problema Zhen propone una estrategia en la que las distancias del centro del ojo derecho e izquierdo (r y l) a la línea que divide el rostro se emplean para calcular el desplazamiento de las rejillas, como se muestra en la Fig. 15. La reubicación de las rejillas a partir de la pose permite deformaciones entre los componentes del rostro (ojos, boca, nariz) y hace al algoritmo más robusto a variaciones en la pose y la expresión. Además, el solapamiento de las rejillas de diferentes componentes hace al método tolerable en cierto grado a errores en la localización de los componentes faciales.

Fig. 15. Extracción de rasgos basados en la pose [9].

En la construcción del índice invertido, luego de ubicar las rejillas en la región del rostro, se extrae de cada región un Histograma de Gradiente (HOG, por sus siglas en inglés) [47]. Luego mediante el modelo *BoW* y el uso del algoritmo de agrupamiento *K-means* jerárquico se entrena el diccionario para la cuantificación de los descriptores. En el proceso de *re-ranking* primero se divide el rostro en parches de 4×4, y luego se obtiene un histograma LBP de cada uno. Los histogramas resultantes se concatenan para obtener un vector de rasgo global de dimensión 944 que representa el rostro. Para un almacenamiento y recuperación eficiente, este vector es reducido mediante PCA a un descriptor G de dimensión 200. Luego a partir del vector G se obtiene una firma hamming, más eficiente que el vector LBP original en cuanto a costo computacional y almacenamiento.

En la experimentación realizada por Zhen [9], se empleó un detector de rostros para obtener $160{,}000$ rostros de la web que se usaron como base de datos básica. Se seleccionan además 40 grupos de imágenes etiquetadas de la base de datos LFW, para tener un total de 685 imágenes de rostros que representan el set de datos del *ground truth*. Como métrica del rendimiento del algoritmo de recuperación se emplea de igual manera la mAP.

El objetivo principal de la experimentación es la evaluación del desempeño del algoritmo atendiendo a diferentes aspectos: (i) la estrategia de rasgos locales basados en la pose, (ii) el método de *re-ranking* multireferencial y (iii) el costo computacional. En cada uno de estos aspectos el algoritmo propuesto tiene buenos resultados, aunque es importante mencionar que no se compara con ningún algoritmo de RIRBC del estado del arte. En los experimentos realizados se compara cada aspecto con diferentes enfoques, no algoritmos. Ejemplo de ello son los experimentos realizados para comprobar el desempeño del algoritmo atendiendo al costo computacional, en este estudio se compara el algoritmo propuesto con una búsqueda lineal.

El algoritmo propuesto por Zhen intenta dar solución a algunos de los problemas identificados en el algoritmo [7] propuesto por Wu, como el proceso de cuantificación supervisado y la variación en la pose. En el proceso de cuantificación propuesto por Zhen es necesario destacar la construcción del vocabulario visual mediante el uso del algoritmo de agrupamiento jerárquico k-*means*. Además Zhen realiza un novedoso aporte para dar solución a la variación de la postura del rostro al proponer la extracción de rasgos locales basados en la pose. Este algoritmo sin embargo al igual que el de Wu requiere tener varias imágenes de una misma persona para hacerle frente a las variaciones intraclases. Además, el empleo *BoW* y *k-means* jerárquico para la construcción del diccionario de palabras visuales tiene como consecuencia que cada rasgo cuantificado solo puede ser asignado a un centroide. Esta restricción se considera que es muy estricta debido a que un rasgo podría ser asignado a más de un centroide. Hay que tener en consideración que la experimentación realizada en este trabajo no es suficiente ya que no se compara siquiera con el algoritmo propuesto por Wu [7], del cual adoptó su metodología.

Recuperación de imágenes de rostro utilizando representación dispersa

La representación dispersa (conocida como ***sparse representation*** ó ***sparse coding*** en inglés) [27] ha evidenciado ser una herramienta extremadamente útil para la adquisición, representación, y compresión de señales de gran dimensión [48]. Como instrumento para modelar señales estáticas, se ha empleado con éxito en aplicaciones para el procesamiento de imágenes, y recientemente ha conducido a buenos resultados en el área del reconocimiento facial [49]. Existen varios trabajos [27, 48-53] vinculados al reconocimiento facial a partir del uso de la representación dispersa (RD).

Recientemente se ha empleado la RD como herramienta para representación a alto nivel de imágenes en sistemas de RIRBC mediante el empleo de estructuras de indexación. A partir del año 2011 Chen [2, 5] ha liderado esta línea de investigación, con el objetivo de desarrollar un sistema escalable de recuperación de imágenes de rostros. En estos trabajos [2, 5] Chen, a partir de las ventajas de la codificación dispersa, propone un sistema de recuperación de rostros empleando un índice invertido construido mediante la RD de las imágenes.

En el primer trabajo realizado [5], Chen identifica como uno de los principales problemas que afecta a los métodos tradicionales de recuperación de imágenes la gran variación (pose, expresión) entre elementos de una misma clase. Además tiene en consideración que en la recuperación de imágenes de rostros es posible tener información extra almacenada en las bases de datos sobre la identidad de una persona, género, etc. Partiendo de esto, propone un algoritmo escalable para la recuperación de imágenes de rostros, que tome en consideración información parcial sobre la identidad de la persona con el objetivo de mejorar los resultados. Para esto propone la RD de rasgos locales extraídos de las imágenes, que luego a partir de un esquema de codificación son refinados mediante la información de la identidad, y empleados en la construcción de un índice invertido. Con el esquema de codificación propuesto, las imágenes de los rostros que presenten una alta variación intraclase serán cuantificadas en palabras visuales similares si poseen la misma identidad como se puede observar en la Fig. 16.

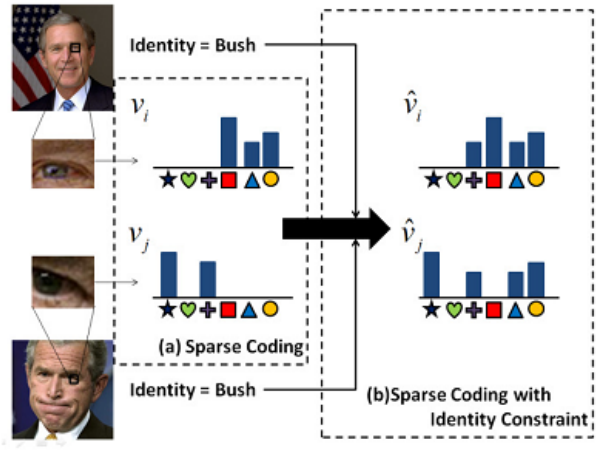

Fig. 16. Algoritmo basado en el empleo de la RD de rasgos locales y el uso de información de identidad [5].

La primera etapa del algoritmo presentado por Chen consiste en la extracción de los rasgos locales y se basa en el algoritmo propuesto por Wu [7]. Para la detección del rostro en las imágenes de la base de datos se emplea el detector en cascada Viola-Jones [54]. La ubicación de los principales componentes del rostro se realiza mediante un Modelo Activo de Forma (conocido como *ASM* por sus siglas en inglés) [55], y la normalización geométrica se realiza mediante la posición de los ojos. Luego se define una rejilla de $5\ X\ 7$ para cada componente del rostro, y se obtiene un total de 175 parches de cada rostro. De cada rejilla se extrae un descriptor LBP [4] de dimensión 59, para conformar un total de 175 descriptores que son cuantificados en palabras visuales separadas mediante la RD. Las imágenes con información de identidad son refinadas mediante el esquema de indexación propuesto, y luego se construye un índice invertido con la representación dispersa final. Al igual que en el algoritmo [7] el descriptor de cada cuadrícula es cuantificado de forma independiente, por lo que dos palabras visuales cotejarán solo si son extraídas del mismo componente facial y las cuadrículas tienen similar ubicación. Esto además permite codificar en las palabras visuales información geométrica de los rostros.

En la etapa de recuperación, a una imagen de consulta se aplica la misma detección de rostro, normalización y extracción de rasgos descrita antes. Los rasgos extraídos son cuantificados en palabras visuales, que a partir de las representaciones almacenadas en el índice invertido y la intersección de histogramas como técnica para calcular similitud, permiten la recuperación de los rostros parecidos a la imagen de consulta. En la Fig. 27 se puede observar un esquema del método propuesto.

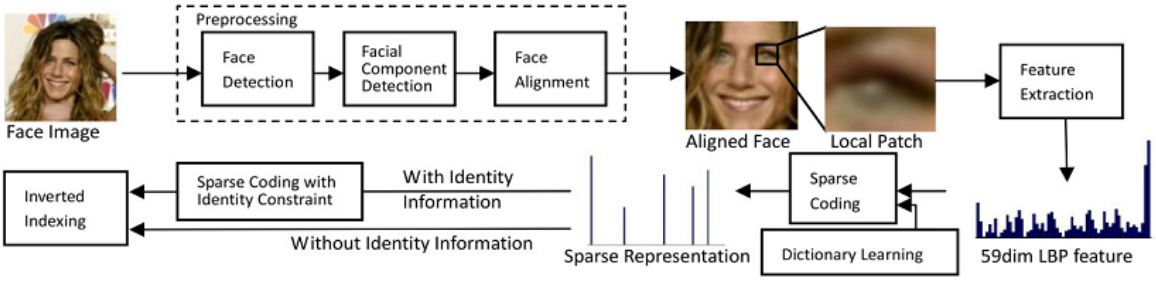

Fig. 17. Esquema del primer algoritmo propuesto por Chen [5].

La razón principal por la cual Chen decide emplear RD en lugar de *BoW* como defiende Wu en [7], es porque en BoW generalmente los autores emplean el algoritmo de agrupamiento $k-means$ para

realizar el entrenamiento del diccionario para la cuantificación [5]. Este algoritmo resuelve el problema de optimización de la Ecuación (5).

$$\min_{D,V} \sum \|x_i - Dv_i\|^2, \qquad (5)$$

$$Card\ (v_i) = 1, \|v_i\|_1 = 1, v_i \geqslant 0, \forall i,$$

donde $D = [d_1, \dots, d_k]$ es la matriz de un diccionario de tamaño $59 \times K$, en el que cada columna representa un centroide (K centroides en total). $V = [v_1, \dots, v_i]$ representa los indicadores de los centroides, cada v_i indica el centroide en D al que pertenece el rasgo x_i. La restricción $Card(v_i) = 1$ significa que cada rasgo x_i solo puede ser asignado a un centroide. Esta restricción se considera que es muy estricta debido a que un rasgo podría ser asignado a más de un centroide, por lo que se sugiere emplear un término regulizador $L1$ en v_i, lo que conlleva a otro problema de optimización conocido como representación dispersa (RD), ver Ecuación (6).

$$\min_{D,V} \sum \|x_i - Dv_i\|^2 + \lambda \|v_i\|_1, \qquad (6)$$

$$sujeto\ a\ \|v_i\| \leq 1, \forall i.$$

Este problema tiene una solución eficiente mediante el algoritmo de optimización *online* basado en aproximaciones estocásticas propuesto por Mairal [56], que es utilizado por Chen para entrenar el diccionario D. Debido a que se cuantifican los descriptores de cada rejilla de forma independiente es necesario entrenar un diccionario diferente para cada uno, por lo que en total son 175 diccionarios a entrenar.

Una vez que se entrena el diccionario D este se puede ajustar en la fórmula anterior y minimizar la función objetivo con v_i para encontrar la RD de cada rasgo. Una vez ajustado D el problema de optimización se convierte en un mínimo cuadrático con regularización $L1$. Debido a que el término de regularización de $L1$ hace a la función objetivo no diferenciable cuando v_i contiene elementos de valor 0, Chen emplea el algoritmo LARS [57] para resolver este problema. La regularización $L1$ conlleva a que cada representación dispersa v_i solo contiene algunos elementos con valor diferente de 0 (conocidos en inglés como *non-zero*) en una dimensión K. Estos elementos *non-zero* son considerados como la palabra visual del descriptor x_i. Debido a que se extraen 175 rasgos, el diccionario general, que contiene las representaciones dispersas v_i de los descriptores de un rostro, tiene una dimensión de $175 \times K$.

Con el objetivo de lograr una menor variación intraclase, Chen propone aplicar un refinamiento a la representación dispersa mediante una restricción de identidad. Luego de este refinamiento, las representaciones dispersas de una misma identidad propagaran las palabras visuales entre ellas. Además, al realizar una consulta se recuperara la mayor cantidad de imágenes de una misma identidad, siempre y cuando al menos una de las imágenes sea similar a la de la consulta. Luego de obtener las RD de los descriptores se realiza el refinamiento mediante un método que aplica una restricción de identidad a cada RD v_i de forma independiente, ver Ecuación (7)

$$\min \beta_{\hat{v}_i} \|x_i - D\hat{v}_i\|^2 + (1-\beta)\|V^p - \hat{v}_i 1^T\|_F^2 + \gamma \|\hat{v}_i\|_1, \qquad (7)$$

donde $V^p = [v_{p1}, \dots, v_{pm}]$ representa las m RD que comparten la misma identidad de v_i, β es un parámetro para ajustar en peso entre la información de identidad y el rasgo visual, y el término γ se emplea para ajustar la dispersión del resultado.

En la experimentación de este primer algoritmo propuesto por Chen, este empleó todas las imágenes de la base de datos LFW. De esta base de datos se escogió como conjunto de prueba 10 imágenes de 12 individuos, un total de 120 imágenes. Como métricas a medir en los experimentos se utilizó la **media del promedio de precisión** (conocido como mAP por sus siglas en inglés) y la **precisión en uno** (p@1) como medida para el rendimiento del algoritmo.

Empleó dos descriptores faciales con búsqueda lineal como líneas bases para comparar el algoritmo propuesto. La primera línea base, denominada (BF), está basada en el descriptor para el reconocimiento facial propuesto en [58]. Se divide el rostro en rejillas de 7×7 y se extraen rasgos LBP de dimensión 59 de cada una, que luego son concatenados para formar un descriptor global de dimensión 2,891. La segunda línea base, denominada (BC) es la concatenación de los descriptores locales extraídos en el algoritmo propuesto. Chen comparó estas líneas bases con el algoritmo propuesto para la codificación dispersa (SC) junto con el método de *re-ranking* (R) empleando la restricción de identidad (I). En la Tabla 1 se resume el rendimiento del método propuesto y las líneas bases, donde se puede observar que el algoritmo propuesto supera a las líneas bases.

Tabla 2. Rendimiento del primer algoritmo propuesto por Chen.

Método	BF	BC	SC	BF+R	BC+R	SC+I
MAP	0.10	0.12	0.16	0.33	0.42	0.72
P@1	0.61	0.71	0.80	0.61	0.71	0.86

Entre los aportes realizados por Chen a la metodología propuesta por Wu resalta el empleo de la representación dispersa (RD) para representar las imágenes de los rostros a un alto nivel. Esta herramienta ha demostrado obtener buenos resultados en el reconocimiento facial, y además su uso ha adquirido auge en la solución de problemas en el área de visión por computadoras en los últimos años [49]. Con el uso del RD Chen además da solución al sobreajuste al que se ve sometido el modelo de *BoW*. Otro aspecto a señalar en este algoritmo es el proceso de codificación basado en la restricción de identidad, con el cual se logra resolver el problema de la variación intraclase. Sin embargo hay que tener en consideración que esta codificación propuesta por Chen es semisupervisada, lo que representa una gran limitación. Esto conlleva a que si la base de datos con la que se desea trabajar no contiene información extra sobre la identidad del rostro, esta tiene que ser proporcionada manualmente. También es necesario considerar que la construcción del diccionario es costosa en cuanto al tiempo, por lo que se podría valorar algunas de las vías propuestas por Wright en [48] para obtener el diccionario. A pesar que este algoritmo resulta un gran paso de avance en la RIRBC la experimentación realizada por el autor quizás no es suficiente.

Los métodos tradicionales para la recuperación de imágenes [34], [5], [7] emplean rasgos de bajo nivel que carecen de significado semántico para representar los rostros. Las imágenes de rostros por lo general presentan una alta variación intraclase (expresión, pose, iluminación), por lo que estos rasgos de bajo nivel afectan los resultados en la recuperación [2]. En esta dirección, en el segundo trabajo realizado [2], Chen da un gran paso de avance al emplear atributos humanos (rasgos débiles) detectados de forma automática que contienen información semántica (género, color de piel, etnia, etc.). Al incorporar atributos humanos de alto nivel a la representación de los rostros y a la estructura de indexación Chen presenta una nueva perspectiva en la RIRBC. Como se muestra en la Fig. 18 imágenes de diferentes individuos pueden representarse de forma similar atendiendo a rasgos de bajo nivel. Al

combinar rasgos de bajo y alto nivel se puede alcanzar una mejor representación y mejorar los resultados en la recuperación. En este trabajo se proponen dos métodos que trabajan en dos direcciones opuestas: (i) **codificación dispersa mejorada a partir de atributos** (conocido como *attribute-enhanced sparse coding* en in inglés) e (ii) **índice invertido mediante la integración de atributos** (conocido como *attribute-embedded inverted indexing* en in inglés). El primer método explota la estructura global del espacio de rasgos y emplea atributos humanos combinados con rasgos de bajo nivel para construir palabras visuales semánticas. El segundo método analiza de forma local los atributos de un rostro representados en un código binario y provee una recuperación eficiente.

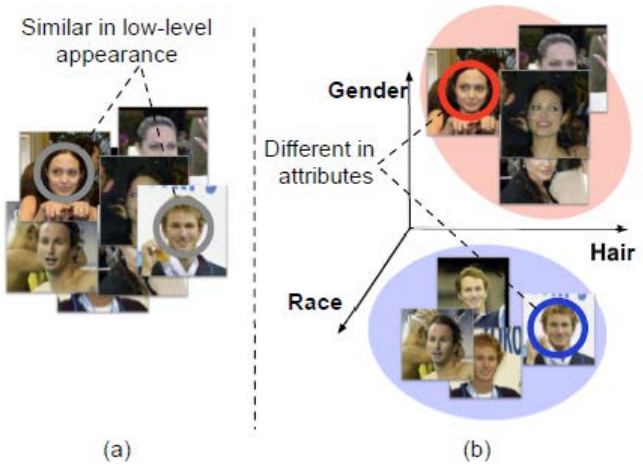

Fig. 18. (a) Grupo de imágenes de rostros representadas mediante rasgos de bajo nivel. (b) Agrupamiento a partir de la incorporación de rasgos débiles [2].

En trabajos como [34], [7] y [5], con el objetivo de reducir la variación intraclase debido a cambios en la pose o iluminación, la región del rostro es recortada y normalizada a una misma posición e iluminación. A causa de este pre-procesamiento se pierde una gran cantidad de información como color de la piel, corte de pelo o género de la persona. Uno de los aspectos más significativos de este segundo trabajo realizado por Chen es la decisión de emplear atributos humanos detectados automáticamente para compensar la pérdida de información debido a la normalización de la imagen del rostro. Además en la perspectiva de la **teoría de la información** el conocimiento de atributos puede reducir la entropía a la hora de identificar una persona. Dada una imagen de un rostro, si X es una variable aleatoria que representa la identidad de una persona y Y el atributo, la información obtenida se puede calcular como:

$$I(X;Y) = H(X) - H(X|Y), \tag{8}$$

donde H(X) es la entropía de Shanon de la variable X, que se emplea para medir la incertidumbre de X. La entropía condicional de X dado Y es H(X|Y), que muestra la incertidumbre de X luego de conocer el valor Y. Mientras mayor sea la información mutua I(X; Y), mayor es la ayuda de Y para predecir X.

En el algoritmo propuesto Chen realiza una primera etapa de pre-procesamiento similar al propuesto en su primer trabajo [5]. Se emplea el mismo método para la detección del rostro en cada imagen de la base de datos. Luego se usa el algoritmo propuesto en [59] para la obtención de 73 indicadores de atributos humanos (género, color del pelo, etnia, etc.). Se reutiliza el *ASM* para localizar 68 marcas

faciales a las que se les aplica un proceso de mapeo basado en coordenadas baricéntricas para alinear el rostro. Se extraen de cada componente facial detectado rejillas de 7×5 para obtener al igual que en el trabajo anterior, un total de 175 descriptores LBP con dimensión 59. Luego de obtener todos los descriptores locales, estos se cuantifican en palabras visuales mediante la codificación dispersa mejorada a partir de atributos. Después se construye el índice invertido mediante la integración de atributos para una recuperación eficiente. Una imagen de consulta es sometida al proceso anterior para obtener la representación dispersa (RD) y los atributos humanos, para luego mediante las palabras visuales y el código binario de los atributos realizar una recuperación de imágenes en el índice.

El proceso de codificación dispersa mejorada a partir de atributos consta de dos etapas: (i) la RD de la imagen y (ii) el mejoramiento de esta representación a partir de los atributos humanos identificados. Es importante mencionar que este proceso se aplica a cada rejilla de una imagen para encontrar 175 palabras visuales que combinadas representan la imagen. El problema de optimización que se resuelve mediante la representación dispersa es el mismo presentado por Chen en [5], en el que la RD de un rasgo no es más que la combinación linear de las columnas del diccionario. Como se puede observar la Ecuación (4) contiene dos partes (i) entrenamiento del diccionario (encontrar D) y (ii) codificación dispersa del rasgo (encontrar V). En [60] Coates demuestra que un diccionario generado a partir de un muestreo aleatorio de rasgos tiene un rendimiento similar a un diccionario entrenado, siempre y cuando los rasgos de la muestra proporcionen un conjunto base capaz de representar una imagen de entrada. Debido que el entrenamiento de 175 diccionarios con una dimensión de 1600 se demora más de dos semanas y teniendo en cuenta los resultados obtenidos por Coates, Chen propone realizar un muestreo aleatorio para generar el diccionario D y pasar directamente a la obtención de V. Luego de obtener D el problema de optimización se convierte en un mínimo cuadrático $L1$ regularizado que se puede resolver con el algoritmo LARS. En la Fig. 19 se puede observar de forma resumida el segundo algoritmo propuesto por Chen.

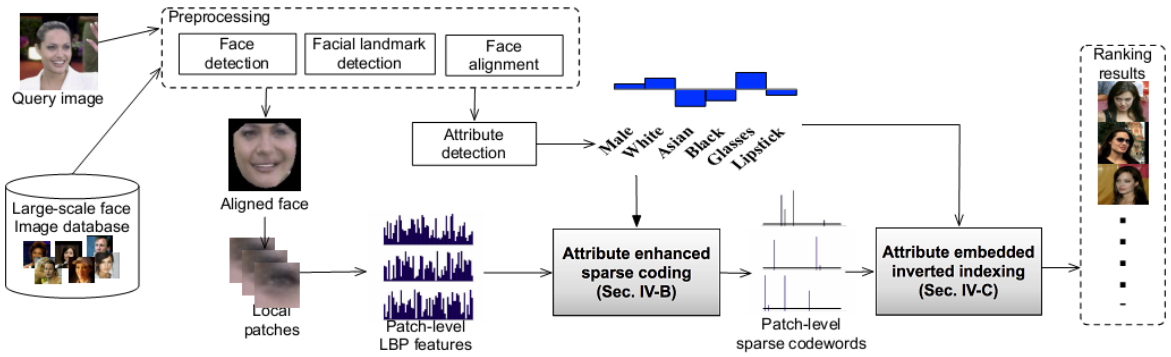

Fig. 19. Esquema del Segundo algoritmo propuesto por Chen [2], que incluye el empleo de atributos humanos.

En el mejoramiento de la RD obtenida se propone dos vías: (i) selección del diccionario D para forzar las imágenes con diferentes valores de atributos a contener diferente codificación dispersa y (ii) asignación de pesos a partir de los valores de atributos. La primera vía consiste en dividir el diccionario en *sub-sets* atendiendo a los valores de atributos. Para una mayor comprensión, atendiendo al atributo de género masculino, el diccionario se dividirá en dos: una parte representa las imágenes que tengan valores positivos del atributo y otra los negativos. En la segunda vía Chen plantea integrar el valor o puntuación obtenido de cada atributo humano a la Ecuación (6) mediante un vector de pesos $z^{(i)}$ que haga la función de una máscara. Basado en [61], Chen primero asigna a la mitad de los centroides del diccionario un valor de atributos $+1$, los que emplea para representar las imágenes que tengan valores de atributos positivos. A la otra mitad de los centroides se le asigna -1 para representar las imágenes con valores de atributos negativos. Luego de la asignación se emplea la distancia entre los valores de

los atributos de una imagen y el valor asignado a los centroides como pesos para seleccionar la palabra visual que represente esa imagen. Debido a que los pesos se obtienen a partir del valor de cada atributo, dos imágenes con valores de atributos similares tendrán un vector de pesos similares, y por tanto tendrán mayor probabilidad de tener una codificación dispersa semejante. De forma detallada, primero se define un vector de atributos $a \in \{1, -1\}^K$, donde a_j contiene el valor de un centroide j, como se muestra a continuación.

$$a_j \begin{cases} +1, & si\ j \geq \left\lfloor \frac{k}{2} \right\rfloor \\ -1, & otro\ caso. \end{cases} \tag{9}$$

Luego, se define el vector $z^{(i)}$ como:

$$z_j^{(i)} = \exp\left(\frac{d\left(f_a(i), a_j\right)}{\sigma}\right), \tag{10}$$

donde $f_a(i)$ es el valor del atributo de la imagen, $d(f_a(i), a_j)$ es la distancia entre $f_a(i)$ y a_j, y σ se emplea para ajustar la descomposición de los pesos. Al final el problema de optimización es representado como se muestra en la Ecuación (11), y se resuelve mediante una versión modificada del algoritmo LARS, ajustando los pesos de acuerdo a $z^{(i)}$.

$$\min_V \sum_{i=1}^{n} \left\| x^{(i)} - D v^{(i)} \right\|_2^2 + \lambda \left\| z^{(i)} \circ v^{(i)} \right\|_1. \tag{11}$$

La construcción del índice invertido mediante la integración de los atributos humanos de una imagen se divide en dos etapas: la primera es el (i) posicionamiento de la imagen y construcción del índice invertido, y la segunda (ii) integración de los atributos al índice invertido. En la primera etapa cada imagen, luego de obtener su RD, es representada mediante un conjunto $c^{(i)}$ de códigos, que se obtiene a partir de los elementos *non-zero* de la RD. La similitud entre dos imágenes entonces se puede calcular como se muestra en la Ecuación (12). El posicionamiento de una imagen de acuerdo a su valor de similitud se puede encontrar eficientemente mediante la estructura de índice invertido.

$$S(i, j) = \left\| c^{(i)} \cap c^{(j)} \right\|. \tag{12}$$

La integración de los atributos humanos al índice invertido consiste en que para cada imagen, además de obtener los códigos $c^{(i)}$ a partir de rasgos de nivel bajo, se emplea una firma binaria d_b para representar un atributo $b^{(i)}$. Debido a esto el valor de similitud se modifica como se puede apreciar en la Ecuación (13).

$$b_j^{(i)} \begin{cases} 1 & si\ f_a^{(i)}(j) > 0 \\ 0 & otro\ caso, \end{cases} \tag{13}$$

$$S(i, j) \begin{cases} \left\| c^{(i)} \cap c^{(j)} \right\| & si\ \ h(b^{(i)}, b^{(j)}) \leq T \\ 0 & otro\ caso. \end{cases} \tag{14}$$

En la Ecuación (14) h(i, j) representa la distancia hamming entre i y j, T es un umbral que cumple la condición $0 \leq T \leq d_b$. El posicionamiento de una imagen se puede obtener de forma eficiente empleando el índice invertido, solo hay que realizar la operación XOR para calcular la distancia hamming antes de actualizar los valores de similitud, como se puede observar en Fig. 20. Debido a que la operación XOR es más rápida que la actualización de los valores de similitud, imágenes con una distancia hamming grande son descartadas, por lo que el tiempo de recuperación disminuye considerablemente.

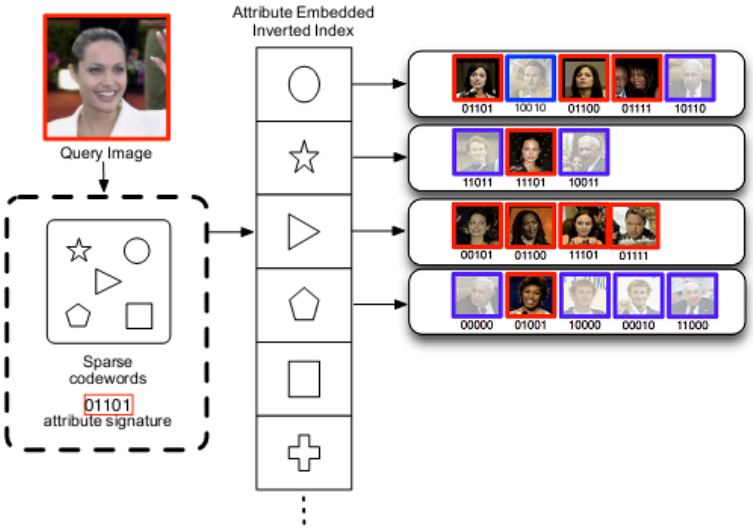

Fig. 20. Esquema del índice invertido con la integración de los atributos faciales [2].

En este segundo algoritmo Chen emplea para la experimentación dos bases de datos públicas: LFW y Pubfig [62]. Como conjunto de prueba utilizó la misma selección de imágenes de LFW que en su primer trabajo. De Pugfig se seleccionó 100 imágenes de 43 individuos (4,300 imágenes) como base de datos para la experimentación, y 10 imágenes por cada uno de los 43 individuos (430 imágenes) como set de prueba.

Chen utiliza diferentes líneas bases para comparar el método propuesto, incluyendo dos descriptores para reconocimiento de rostros del estado del arte. Estos métodos son: (1) LBP: consiste en concatenar descriptores LBP [58] uniformes de dimensión 59, obtenidos de 1os 175 parches antes descritos, para formar un vector de dimensión 10325; (2) ATTR: obtención de los 73 valores de atributos humanos mediante el algoritmo propuesto en [59]; (3) SC: representación dispersa obtenida de descriptores LBP mediante el muestreo aleatorio de 1600 muestras como diccionario de centroides y combinado con índice invertido; (4) SC-original: similar al método (3), pero empleando de forma directa el peso de la RD en una búsqueda lineal, en lugar de emplear índice invertido; (5) ASC-D: codificación dispersa mejorada a partir de atributos usando la vía de selección de diccionario; (6) ASC-W: codificación dispersa mejorada a partir de atributos usando la asignación de pesos; (7) AEI: construcción índice invertido mediante la integración de atributos.

Como métricas a medir en los experimentos Chen emplea la mAP y la **precisión en K** (p@*X*). Como paso inicial para una correcta experimentación se configuró el tamaño del diccionario $K = 1600$ y el valor de $\lambda = [10^{-6}; 10^{-2}]$ para la RD. Además para el caso del método ASC-W se fijó $\sigma = 120$. En la experimentación realizada Chen analiza el rendimiento de las líneas bases y de los métodos ASC-W,

ASC-D y AEI atendiendo a diferentes aspectos; e identifica los atributos que brindan más información sobre la identidad de la persona para cada base de datos. Además mide el rendimiento del algoritmo propuesto atendiendo a la escalabilidad y se muestran dos ejemplos de prueba empleando SC y ASC-W respectivamente.

En la Tabla 3 se muestran los resultados obtenidos por Chen al comparar las líneas bases LBP, ATTR, SC y SC-original, en donde resalta el buen desempeño del método LBP por ser más robusto a variaciones de pose. Los resultados de SC y ATTR sugieren que la codificación dispersa mejorada a partir de atributos podría mejorar la eficiencia de los sistemas de RIRBC. Además se puede observar en los resultados de SC y SC-original la superioridad del índice invertido en todas las métricas, sobre todo en el tiempo de recuperación.

Tabla 3. Rendimiento del segundo algoritmo propuesto.

Base de datos	LFW			Pubfig		
# de personas	5749			43		
Tamaño de la base de datos	13113			4300		
# de consultas	120			430		
Rendimiento	MAP	P@10	Time(s)	MAP	P@10	Time(s)
LBP	11.9%	49.6%	1.01	11.6%	47.4%	0.38
ATTR	11.6%	37.8%	0.04	15.1%	39.7%	0.01
SC	13.0%	46.8%	0.03	14.7%	49.0%	0.01
SC-original	10.0%	39.3%	1.73	10.9%	39.6%	0.59

En la Fig. 21 se muestra el rendimiento del método AEI en la base de datos Pubfig empleando diferentes valores para el umbral T (ver Ecuación (14)). Cuando a T se le asigna un valor elevado, el rendimiento converge en SC, debido que se ignora el código de los atributos. Sin embargo cuando el valor de T es pequeño el rendimiento mejora, pero si es demasiado pequeño este afecta de forma el rendimiento. Según explica Chen, existen dos posibles motivos por los cuales se afecta el rendimiento: (i) error en la detección de los atributos o que (ii) algunos de los atributos detectados no son efectivos en la identificación de un individuo. Para verificar la segunda posible causa, el autor realizó los mismos experimentos teniendo en consideración solo los primeros 40 atributos del *ranking* obtenido en los resultados del método ASC-W antes mencionado. Al descartar atributos que no brindan información sobre la identidad de la persona se logra superar el rendimiento.

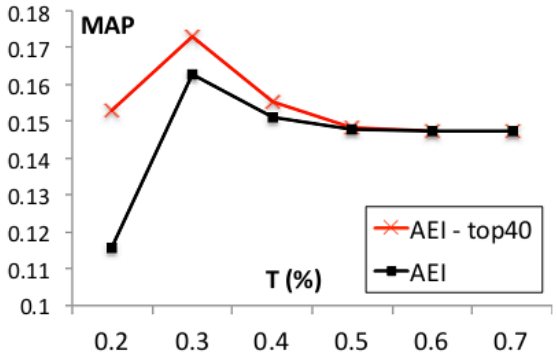

Fig. 21. Resultados del procedimiento AEI [2].

La combinación de los métodos ASC-W y AEI permitió mejorar el rendimiento en la recuperación. En la Tabla 4 se muestran los resultados obtenidos al combinar estos métodos, en los que resalta la mejora relativa de 43.55% y 42.39% sobre SC. Aunque esta mejora se debe fundamentalmente a ASC-W, AEI es convertido de forma exitosa en un índice invertido y es capaz de reducir el tiempo de recuperación. Además se puede observar cómo se mejoran los resultados en el mAP y la precisión en la selección de los mejores candidatos (P@10 y P@20).

Tabla 4. Resultados al combinar ASC-W y AEI.

LFW	MAP	Mejora Relativa	Mejora Absoluta	P@10	P@20
SC	13.0%	-	-	46.8%	36.2%
ASC-W	18.3%	41.0%	5.3%	57.2%	45.3%
AEI	14.3%	10.5%	1.3%	49.0%	37.5%
ASC-W+AEI	18.6%	43.6%	5.6%	57.3%	45.5%
Pubfig	MAP	Mejora Relativa	Mejora Absoluta	P@10	P@20
SC	14.8%	-	-	49.0%	40.4%
ASC-W	20.2%	36.9%	5.4%	56.4%	47.7%
AEI	17.6%	19.3%	2.8%	51.4%	42.8%
ASC-W+AEI	21.0%	42.4%	6.2%	56.9%	48.3%

En el segundo algoritmo [2] Chen da solución a algunos problemas detectados en el método [5] y realiza un importante aporte al emplear rasgos débiles para mejorar la RD y la recuperación de imágenes de rostros. Uno de los problemas que se solucionan es el costoso (atendiendo al tiempo) entrenamiento del diccionario, en este algoritmo se propone construir el diccionario a partir de un muestreo aleatorio que brinda similares resultados sin un costo de tiempo tan elevado. Sin embargo es válido mencionar que se podrían tener en cuenta y analizar otra de las opciones propuestas por Wright en [48] para construir el diccionario. Otros aspectos relevantes a mencionar de este algoritmo es la ejecución de dos procesos que tienen gran influencia en los resultados obtenidos: (i) la codificación dispersa mejorada a partir de atributos faciales usando la asignación de pesos y (ii) la construcción de un índice invertido mediante la integración de atributos faciales. El primer proceso tiene suma importancia ya que mejora las representaciones dispersas, soluciona los problemas de variación intraclase y resuelve el inconveniente del algoritmo [5] de ser semisupervisado. Es importante señalar

que a pesar del margen de mejora obtenido con el empleo de atributos faciales, se debería realizar un estudio más profundo sobre los tipos de rasgos que se pueden extraer las diferentes regiones de interés de un rostro para la descripción del mismo según sus características faciales. Analizar qué información es la más adecuada extraer de los píxeles (intensidad, RGB, HSV, etc.) de acuerdo con los atributos que son relevantes en la identidad de una persona. Además estudiar si es necesario normalizar y seleccionar el método más eficaz, y también se podría considerar que tipo de estadística se obtendrá como resultado (variancia, histograma, etc.). El segundo proceso tiene gran impacto en la velocidad de recuperación ya que al incorporar una firma a partir de los atributos detectados se mejora el proceso de *ranking*. Esto se debe a que se reduce el espacio de búsqueda cuando se desechan imágenes que presentan firmas de atributos diferentes. En la etapa de experimentación Chen demuestra que no todos los rasgos débiles extraídos de un rostro brindan información relevante sobre la identidad de una persona, un aporte importante a tener en cuenta para un análisis más consciente en la selección de estos atributos. En resumen, este algoritmo presenta claras ventajas sobre los métodos anteriores, pero presenta un esquema general con aspectos que pueden ser mejorados.

4 Comparación de los sistemas basados en apariencia local

Como se analizó en la sección 3, los algoritmos que emplean rasgos basados en forma [28] presentan tres limitantes fundamentales: (i) no brindan suficiente información sobre la imagen de un rostro, (ii) no tienen en cuenta la apariencia y (iii) necesitan la localización exacta de los puntos característicos del rostro. Por otra parte, los algoritmos basados en rasgos de apariencia global [6, 34] se ven seriamente afectados por variaciones de pose e iluminación, y son más sensibles a errores en la detección y alineación de los rostros en las imágenes [4]. Por esta razón, en esta sección se hace un resumen comparativo entre los principales algoritmos basados en rasgos de apariencia local analizados.

La experimentación realizada por los autores de los algoritmos analizados en el estado del arte es insuficiente y superficial, debido a que no se realizan comparaciones con otros métodos de RIRBC bajo las mismas condiciones. La experimentación por lo general es pobre, se analizan solamente resultados basados en variaciones del mismo algoritmo, atendiendo a diferentes descriptores, o diferentes parámetros de configuración. Esto se debe tal vez a que esta área de investigación no ha sido muy explorada, no existe un resumen del estado de arte de los sistemas de RIRBC y a que algunos autores no han publicado sus metodologías y configuraciones para la experimentación, lo que hace imposible compararse con ellos.

Los algoritmos basados en apariencia local analizados en este trabajo proponen de forma general el empleo de estructuras de indexación partiendo de descriptores locales y de la información geométrica para representar las imágenes de los rostros. En el caso del algoritmo propuesto por Kaushik [40] se emplea el descriptor SURF para localizar puntos de interés en el rostro y extraer sus vectores de rasgos y coordenadas, información que se almacena en una tabla hash mediante el método *geometric hashing*, que garantiza la conservación de la información espacial de estos puntos. Este método resulta interesante ya que emplea información de apariencia y espacial para describir un rostro, aunque tiene algunos puntos a considerar como es el uso del descriptor SURF para la extracción de los rasgos, el cual no tiene en cuenta las características faciales que describen un rostro (boca, nariz y ojos). Además, a pesar de ser adecuado para la tarea de reconocimiento facial según Geng Du [42], no se han realizado estudios validen su eficacia respecto los descriptores más utilizados en el estado del arte.

Los algoritmos propuestos por Wu [7] y Zhen [9] emplean una metodología que se basa en la extracción de rasgos locales de los principales componentes que describen un rostro (ojos, boca y nariz) y la construcción de un diccionario de palabras visuales a partir de estos rasgos. Estos métodos tienen

en cuenta la posición de los rasgos extraídos y extraen rasgos globales para realizar un proceso de *reranking* que mejora la recuperación de candidatos. Si se tiene en consideración la metodología propuesta por estos algoritmos, se puede concluir que estos superan al método propuesto por Kaushik. Además se puede decir que el algoritmo propuesto por Zhen supera al presentado por Wu, ya que este brinda una vía para dar solución a los problemas de pose y propone la construcción del diccionario de palabras visuales de forma automática. Sin embargo hay que recalcar que los resultados obtenidos en estos trabajos no son comparables ya que en la actualidad no existe un protocolo para la experimentación que lo permita, la obtención y construcción de los conjuntos de datos es diferente para cada algoritmo, las bases de datos empleadas no son las mismas y algunas no son públicas. Además los parámetros medidos por cada autor no son los mismos, y no se ha definido una línea base con la cual compararse.

Los algoritmos presentados por Chen [2, 5] siguen la metodología propuesta por Wu. Estos mejoran los métodos propuesto por Wu y Zhen ya que proponen el uso de RD para para representar las imágenes de los rostros a un alto nivel. El empleo de RD da solución al sobreajuste al que se ven sometidos los modelos de *BoW* propuestos por Wu y Zhen, una desventaja importante que presentan estos enfoques. Además esta herramienta tiene buenos resultados en el reconocimiento facial y ha adquirido auge en la solución de problemas en el área de visión por computadoras. Los algoritmos presentados por Chen además solucionan la semi-supervisación a la que se ve sometido el algoritmo de Wu en cuanto a la construcción del diccionario de palabras visuales. Por otra parte hay que destacar el aporte realizado por Zhen en la reubicación de las rejillas sobre los componentes del rostro, con lo que se obtiene más robustez ante los problemas de variación en la pose y la expresión. Atendiendo a las desventajas de los algoritmos propuestos por Wu y Zhen, el auge y los buenos resultados alcanzados por la RD se puede plantear que la mejor línea de investigación a seguir es la desarrollada por Chen.

Utilizando las bases de datos de imágenes de rostro existentes se debe establecer un protocolo de experimentación para comparar los diferentes algoritmos. Las bases de datos empleadas para la experimentación y entrenamiento de los algoritmos basados en rasgos locales son FERET, LFW y PubFig. En la Tabla 5 se muestra varias características de las bases de datos, donde se puede observar que LFW y Pubfig parten con ventaja sobre FERET, debido a que las imágenes son recopiladas de ambientes no controlados, lo que permite simular problemas de iluminación, pose, calidad de la imagen, entre otros que afectan a los algoritmos de reconocimiento facial. Además en la metodología propuesta por Chen en [2], que se considera una de las más completa en la literatura actual, se emplea LFW y Pubfig por lo antes descrito. Por eso se plantea que el protocolo de experimentación debería partir del empleo de las bases de datos LFW y PubFig.

Tabla 5. Bases de datos empleadas para la experimentación y entrenamiento de los algoritmos basados en rasgos locales.

BD	Público	Origen	Controlada	Rostros/ID	IDs conocidos	# Rostros
FERET	si	Laboratorio	si	12	1 199	14 126
LFW	si	Web	no	3	5 749	13 233
PubFig	si	Web	no	300	200	58 797

5 Conclusiones

El estudio del arte realizado sobre los algoritmos de RIRBC muestra que esta área no ha sido del todo explotada, existen relativamente pocos trabajos que enfrentan este problema a pesar de la importancia que tiene para los sistemas de reconocimiento facial. Se ha expuesto que la reducción de espacio de

búsqueda es un problema de gran relevancia que no se puede solucionar simplemente con la reducción de los rasgos extraídos. Un algoritmo eficaz debe implementar una metodología que permita considerar la mayor cantidad de información contenida en la imagen de un rostro y debe emplear una estructura de indexación que permita almacenar esta información de forma eficiente para así garantizar escalabilidad en el proceso de recuperación. En este estudio se evidencia que el empleo de rasgos locales y globales tiene gran relevancia para los procesos de representación y recuperación de rostros. Además se demuestra que el empleo de información geométrica, de conjunto con la información de apariencia, brinda mayor precisión y permite lidiar con problemas de variación intraclase.

A partir del análisis y comparación de los algoritmos estudiados, se puede concluir que la metodología propuesta por Wu y perfeccionada luego por Chen representa un avance significativo en esta área, definiendo un esquema general que debe ser tenido en cuenta a la hora de desarrollar un nuevo sistema de RIRBC. La representación *sparse* propuesta por Chen figura como la mejor metodología del estado del arte para dar solución a la RIRBC. En esta se tiene en cuenta prácticamente toda la información que se puede extraer de la imagen de un rostro para lograr el reconocimiento facial y se da solución a gran parte de los problemas que afectan el proceso de recuperación. Además se realiza un gran aporte al emplear los rasgos débiles (*soft biometrics*) para reducir el espacio de búsqueda. Este paradigma no es una solución definitiva a la RIRBC y se debe considerar como un punto de partida al enfrentar el problema. Esta metodología fue propuesta recientemente, por lo que existen interrogantes y caminos a seguir, existiendo aún un amplio margen de mejora. La información a emplear de la imagen para la extracción de atributos faciales, la selección de estos rasgos débiles, el modo de entrenamiento del diccionario de palabras visuales, el proceso de recuperación de rostros candidatos e incluso el descriptor empleado para la extracción tanto de rasgos locales como globales son algunos ejemplos de tópicos que aún se pueden investigar y perfeccionar.

Referencias bibliográficas

1. Lei, Y.-H., et al., Photo search by face positions and facial attributes on touch devices, in Proceedings of the 19th ACM international conference on Multimedia, 2011, ACM: Scottsdale, Arizona, USA. p. 651-654.
2. Chen, B., et al., Scalable Face Image Retrieval using Attribute-Enhanced Sparse Codewords. Multimedia, IEEE Transactions on, 2013. **PP**(99): p. 1-1.
3. (UIDAI), U.I.A.o.I., Role of Biometric Technology in Aadhaar Authentication, 2012, Planning Commission,Govt. of India (GoI): 3rd Floor, Tower II,Jeevan Bharati Building,Connaught Circus,New Delhi 110001.
4. Zhao, W. and R. Chellappa, Face Processing: Advanced Modeling and Methods, 2006: Elsevier Science.
5. Chen, B.-C., et al., Semi-supervised face image retrieval using sparse coding with identity constraint, in Proceedings of the 19th ACM international conference on Multimedia, 2011, ACM: Scottsdale, Arizona, USA. p. 1369-1372.
6. Kwan-Ho, L., et al. An efficient human face indexing scheme using eigenfaces. in Neural Networks and Signal Processing, 2003. Proceedings of the 2003 International Conference on. 2003.
7. Zhong, W., Scalable Face Image Retrieval with Identity-Based Quantization and Multireference Reranking. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011. **33**(10): p. 1991-2001.
8. Eakins, J.P. and M.E. Graham, Content-based Image Retrieval, 1999, Joint Information Systems Committee.
9. Haiyang Zhen, L.Z., Rong Zhan, Dong Yin, Rapid Face Image Retrieval with Pose-based Local Features and Multi-reference Re-ranking, in Journal of Computational Information Systems, 2013.
10. Sridharan, K. and S.U.o.N.Y.a. Buffalo, Semantic Face Retrieval, 2006: State University of New York at Buffalo.
11. Venkat N. Gudivada, V.V.R., and Guna S. Seetharaman, An Approach to Interactive Retrieval in Face Image Databases Based on Semantic Attributes, in Third Annual Symposium on Document Analysis and Information Retrieval: Proceedings : April 11-13, 1994, Las Vegas, Nevada, 1994, Information Science Research Institute, University of Nevada.

12. Shen, B.-C., C.-S. Chen, and H.-H. Hsu. Face image retrieval by using Haar features. in Pattern Recognition, 2008. ICPR 2008. 19th International Conference on. 2008.
13. Mikolajczyk, K. and C. Schmid, A performance evaluation of local descriptors. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2005. **27**(10): p. 1615-1630.
14. Kafai, M., K. Eshghi, and B. Bhanu, Discrete Cosine Transform Locality-Sensitive Hashes for Face Retrieval. Multimedia, IEEE Transactions on, 2014. **16**(4): p. 1090-1103.
15. Oravec, M., Face Recognition, 2010: In-Tech.
16. Ajit Kumar Mahapatra, S.B., Inverted indexes: Types and techniques. IJCSI International Journal of Computer Science Issues, 2011. **8**(4).
17. Paulevé, L., H. Jégou, and L. Amsaleg, Locality sensitive hashing: a comparison of hash function types and querying mechanisms. Pattern Recognition Letters, 2010. **31**(11): p. 1348-1358.
18. Wolfson, H.J. and I. Rigoutsos, Geometric hashing: an overview. Computational Science & Engineering, IEEE, 1997. **4**(4): p. 10-21.
19. Ahonen, T., et al. Recognition of blurred faces using Local Phase Quantization. in Pattern Recognition, 2008. ICPR 2008. 19th International Conference on. 2008.
20. Winder, S.A.J. and M. Brown. Learning Local Image Descriptors. in Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on. 2007.
21. Moore, A., An introductory tutorial on kd-trees, 1991, Carnegie Mellon University, Pittsburgh, Technical Report.
22. Vleugels, J. and R. Veltkamp. Efficient Image Retrieval through Vantage Objects. in Visual Information and Information Systems. 1999.
23. Li, S.Z. and A.K. Jain, Handbook of Face Recognition, 2011: Springer London, Limited.
24. Tan, X., et al., Face recognition from a single image per person: A survey. Pattern Recognition, 2006. **39**(9): p. 1725-1745.
25. Zhao, W., et al., Face recognition: A literature survey. ACM Comput. Surv., 2003. **35**(4): p. 399-458.
26. O'Hara, S. and B.A. Draper, Introduction to the Bag of Features Paradigm for Image Classification and Retrieval. CoRR, 2011. **abs/1101.3354**.
27. Wright, J., et al., Robust Face Recognition via Sparse Representation. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2009. **31**(2): p. 210-227.
28. Vikram, T.N., et al. Face Indexing and Retrieval by Spatial Similarity. in Image and Signal Processing, 2008. CISP '08. Congress on. 2008.
29. Jesorsky, O., K.J. Kirchberg, and R. Frischholz, Robust Face Detection Using the Hausdorff Distance, in Proceedings of the Third International Conference on Audio- and Video-Based Biometric Person Authentication, 2001, Springer-Verlag. p. 90-95.
30. Samaria, F.S. and A.C. Harter. Parameterisation of a stochastic model for human face identification. in Applications of Computer Vision, 1994., Proceedings of the Second IEEE Workshop on. 1994.
31. Gudivada, V.N., Θℜ-string: A geometry-based representation for efficient and effective retrieval of images by spatial similarity. Knowledge and Data Engineering, IEEE Transactions on, 1998. **10**(3): p. 504-512.
32. El-Kwae, E.A. and M.R. Kabuka, Efficient content-based indexing of large image databases. ACM Trans. Inf. Syst., 2000. **18**(2): p. 171-210.
33. Sciascio, E.D., et al., Retrieval by spatial similarity: an algorithm and a comparative evaluation. Pattern Recogn. Lett., 2004. **25**(14): p. 1633-1645.
34. Wang, D., et al., Retrieval-based face annotation by weak label regularized local coordinate coding, in Proceedings of the 19th ACM international conference on Multimedia, 2011, ACM: Scottsdale, Arizona, USA. p. 353-362.
35. Siagian, C. and L. Itti, Rapid Biologically-Inspired Scene Classification Using Features Shared with Visual Attention. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2007. **29**(2): p. 300-312.
36. A.M. Martinez, R.B., The AR Face Database, in CVC Technical Report, 1998.
37. Graham, D.B.A., N. M., Characterizing Virtual Eigensignatures for General Purpose Face Recognition, in NATO ASI SERIES F COMPUTER AND SYSTEMS SCIENCES; 163; 446-456; Face recognition: from theory to applications. 1998, Springer: NATO Advanced Study Institute.
38. University of Bern face database.; Available from: ftp://iamftp.unibe.ch/pub/Images/FaceImages/.

39. Belhumeur, P.N., J.P. Hespanha, and D. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 1997. **19**(7): p. 711-720.
40. Kaushik, V.D., et al., An efficient indexing scheme for face database using modified geometric hashing. Neurocomputing, 2012(0).
41. Bay, H., et al., Speeded-Up Robust Features (SURF). Comput. Vis. Image Underst., 2008. **110**(3): p. 346-359.
42. Geng Du, F.S., Anni Cai. Face recognition using SURF features. in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. 2009.
43. Phillips, P.J., et al., The FERET Evaluation Methodology for Face-Recognition Algorithms. IEEE Trans. Pattern Anal. Mach. Intell., 2000. **22**(10): p. 1090-1104.
44. Freeman, W.T. and E.H. Adelson, The design and use of steerable filters. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 1991. **13**(9): p. 891-906.
45. Zhimin, C., et al. Face recognition with learning-based descriptor. in Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. 2010.
46. Huang, G.B., et al., Labeled faces in the wild: A database for studying face recognition in unconstrained environments. University of Massachusetts, Amherst, 2007.
47. Dalal, N. and B. Triggs. Histograms of oriented gradients for human detection. in Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. 2005.
48. Wright, J., et al., Sparse Representation for Computer Vision and Pattern Recognition. Proceedings of the IEEE, 2010. **98**(6): p. 1031-1044.
49. Meng, Y., D. Zhang, and Y. Jian. Robust sparse coding for face recognition. in Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. 2011.
50. Gao, S., I.W.-H. Tsang, and L.-T. Chia, Kernel sparse representation for image classification and face recognition, in Proceedings of the 11th European conference on Computer vision: Part IV, 2010, Springer-Verlag: Heraklion, Crete, Greece. p. 1-14.
51. Huimin, G., et al. Face verification using sparse representations. in Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on. 2012.
52. Jiang, Z., Z. Lin, and L. Davis. Learning a Discriminative Dictionary for Sparse Coding via Label Consistent K-SVD. in CVPR. 2011.
53. Qiang, Z. and L. Baoxin. Discriminative K-SVD for dictionary learning in face recognition. in Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. 2010.
54. Viola, P. and M. Jones. Rapid object detection using a boosted cascade of simple features. in Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on. 2001.
55. Milborrow, S. and F. Nicolls, Locating Facial Features with an Extended Active Shape Model, in Proceedings of the 10th European Conference on Computer Vision: Part IV, 2008, Springer-Verlag: Marseille, France. p. 504-513.
56. Mairal, J., et al., Online dictionary learning for sparse coding, in Proceedings of the 26th Annual International Conference on Machine Learning, 2009, ACM: Montreal, Quebec, Canada. p. 689-696.
57. Efron, B., et al., Least angle regression. The Annals of Statistics, 2004. **32**(2): p. 407-499.
58. Ahonen, T., A. Hadid, and M. Pietikäinen, Face Recognition with Local Binary Patterns, in Computer Vision - ECCV 2004, T. Pajdla and J. Matas, Editors. 2004, Springer Berlin Heidelberg. p. 469-481.
59. Kumar, N., et al., Describable Visual Attributes for Face Verification and Image Search. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2011. **33**(10): p. 1962-1977.
60. Coates, A. and A.Y. Ng, The Importance of Encoding Versus Training with Sparse Coding and Vector Quantization, in ICML, L. Getoor and T. Scheffer, Editors. 2011, Omnipress. p. 921-928.
61. Jinjun, W., et al. Locality-constrained Linear Coding for image classification. in Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. 2010.
62. Kumar, N., et al. Attribute and simile classifiers for face verification. in Computer Vision, 2009 IEEE 12th International Conference on. 2009.

RT_063, septiembre 2014
Aprobado por el Consejo Científico CENATAV

Editor: Lic. Lucía González Bayona
Diseño de Portada: Di. Alejandro Pérez Abraham
RNPS No. 2142
ISSN 2072-6287
Indicaciones para los Autores:
Seguir la plantilla que aparece en www.cenatav.co.cu
C E N A T A V
7ma. A No. 21406 e/214 y 216, Rpto. Siboney, Playa;
La Habana. Cuba. C.P. 12200
Impreso en Cuba