UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

CLASIFICACIÓN DE GÉNERO EN IMÁGENES FACIALES USANDO INFORMACIÓN MUTUA

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN CIENCIAS DE LA INGENIERÍA, MENCIÓN ELÉCTRICA

JUAN EDUARDO TAPIA FARIAS

PROFESOR GUÍA:

Dr. CLAUDIO PÉREZ FLORES

MIEMBROS DE LA COMISIÓN:

Dr. MARCOS DÍAZ QUEZADA

Dr. PABLO ESTÉVEZ VALENCIA

Dr. JAVIER RUIZ DEL SOLAR SAN MARTÍN

Dr. PABLO ZEGERS FERNÁNDEZ

SANTIAGO DE CHILE

ABRIL 2012

RESUMEN DE LA TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN CIENCIAS DE LA INGENIERÍA, MENCIÓN INGENIERÍA ELÉCTRICA POR:
JUAN EDUARDO TAPIA FARIAS
FECHA: 04/04/2012
PROF. GUÍA: CLAUDIO PÉREZ FLORES.

Resumen

Durante la década de los 90, uno de los principales problemas abordados en el área de visión computacional fue el detectar rostros en imágenes, para lo cual se desarrollaron innumerables métodos y aplicaciones que pudieran realizar dicha tarea. En la actualidad, ese problema se encuentra prácticamente solucionado con detectores con tasas de detección muy altas, por lo cual, el problema ha evolucionado a poder obtener información adicional de estos rostros detectados, ya sea identificando su raza, edad, emociones, género, entre otros. Es en este contexto, que se enmarca esta investigación.

La clasificación de género se considera una tarea difícil y complementaria al reconocimiento de patrones, a causa de la alta variabilidad de la apariencia del rostro. Los rostros son objetos no rígidos y dinámicos con una diversidad grande en la forma, el color y la textura, debido a múltiples factores como la pose de la cabeza, iluminación, expresiones faciales y otras características faciales. La alta variabilidad en la apariencia de los rostros afectan directamente su detección y clasificación.

En este trabajo de tesis se implementaron los métodos de extracción de características basados en intensidad y textura, se midió su desempeño con 4 tipos de clasificadores distintos. Las características extraídas fueron fusionadas al nivel de las características.

Por otra parte, se extendió el efecto de seleccionar características utilizando 3 métodos basados en Información Mutua, Mínima redundancia y Máxima relevancia(mRMR), Información Mutua Normalizada (NMIFS), Información Mutua Condicional (CMIFS). Se compararon nuestros resultados con los mejores datos publicados, utilizando las bases de datos internacionales de rostros FERET y WEB, usando diferentes tamaños de imágenes y particiones de datos.

Se obtuvieron mejoras significativas en la clasificación de género, que van desde 1.2 % al 12.7 % sobre la base de datos FERET y desde 4.1 % al 8.9 % sobre la base de datos WEB. Además, se redujo el número de características utilizadas como entradas en el clasificador. Dependiendo del tamaño de la imagen, el número total de características seleccionadas es reducida a menos del 74 % en la base de datos FERET y en un 76.04 % en la base de datos

WEB. Por lo tanto, el tiempo computacional se reduce significativamente para aplicaciones en tiempo real.

Abstract

During the 90's, one of the main issues addressed in the area of computer vision was to detect faces in images, which were developed for many methods and applications that could perform the task. At present, this problem is practically solved with very high detection rates, so the problem has evolved to obtain additional information from these detected faces, either by identifying their race, age, emotions, gender, among other. In this context, we develop this this research.

Gender classification is considered a difficult and complementary task to pattern recognition, because of the high variability of the appearance of the face. The faces are non-rigid and dynamic objects with a large diversity in form, color and texture, due to multiple factors such as head pose, lighting, facial expressions and other facial features. The high variability in the appearance of faces directly affect their detection and classification.

In this thesis feature extraction methods were implemented based on intensity and texture and their performance was measured with 4 different types of classifiers. The extracted features were merged at the level of features.

On the other hand, the research also studied the effect of selecting features using 3 methods based on Mutual Information, Minimum Redundancy and Maximum Relevance (mRMR), Normalized Mutual Information (NMIFS) Conditional Mutual Information (CMIFS). We compared our results with the best published data, using international face databases (WEB and FERET), using different sizes of images and data partitions.

We obtained significant improvements in gender classification, ranging from 1.2 % to 12.7 % on the FERET database and from 4.1 % to 8.9 % on the WEB database. In addition, reduced the number of features used as inputs to the classifier. Depending on the size of the image, the total number of selected features is reduced to less than 74 % on FERET database and 76.04 % in the WEB database. Therefore, the computational time is greatly reduced for real time applications.

A mis Padres y Familia,
gracias por el apoyo y la confianza...

Agradecimientos

En primer lugar, quisiera agradecer a quienes son las personas más importantes en mi vida: mi esposa Simone y mi hijo Vinicius, por su infinito apoyo durante mis estudios. Gracias por todos los sacrificios que han realizado para poder brindarme la oportunidad de obtener esta carrera, ya que cada hora de estudio es una hora sin ustedes.

A mis Padres, especialmente a mi madre Patricia por toda su abnegación y sacrificio durante su corta vida, que sin duda estarías muy orgullosa, descansa en paz.

A mis amigos de la U, por los muy gratos momentos vividos dentro y fuera de la Facultad. A mis compañeros del laboratorio, por todas las anécdotas, enseñanzas y momentos alegres entregados durante el desarrollo de mi tesis.

Doy las gracias a mi profesor guía, Claudio Pérez, por sus consejos y darme la oportunidad de trabajar con él en esta tesis.

Esta tesis ha sido parcialmente financiada por los proyectos Fondef D08I1060 y Fondecyt 1120613.

Índice general

Índice de figuras

Capítulo 1

Introducción

1.1. Motivación.

La clasificación automática de género tiene una amplia gama de posibles aplicaciones, en mejorar la interacción hombre máquina y los métodos de identificación de rostro [1–3]. La clasificación de género podría ser usada para particionar bases de datos de reconocimiento de rostros, construir n clústers para reducir el número de comparaciones y así identificar un rostro [4–6]. Otras novedosas aplicaciones, incluyen recolección de información demográfica, evaluación de comportamiento de consumidores y marketing electrónico selectivo en tiendas de retail.

El estudio de la clasificación de género es un tópico emergente comparado con otros métodos de identificación biométrica, tales como huellas dactilares, reconocimiento facial o identificación de iris [7–9]. Una completa revisión bibliográfica y comparación de los mejores métodos de clasificación de género fue realizada en [3, 10]. Se concluyó que había un número muy pequeño de publicaciones con resultados probados y disponibles en bases de datos internacionales para ser replicados.

La mayoría de los artículos comparan diferentes enfoques en la clasificación de género, sin identificar las particiones de las bases de datos, imágenes utilizadas o ejemplos no homogéneos en cuanto a su tamaño. Debido a esto, no es posible replicar los resultados para poder compararse [11–13]. Por lo tanto la revisión bibliográfica que se presenta a continuación se centra sólo en aquellos métodos que pueden ser comparados y con bases de datos estándares.

Cuatro diferentes algoritmos se utilizaron en [10] para la clasificación de género, los cuales fueron comparados utilizando 2 bases de datos internacionales: FERET [14] y WEB [10]. Los métodos utilizados fueron: Una red neuronal multicapa (NN), Máquinas de Soporte Vectorial (SVM) con una matriz de pixeles como base de entrada, diferentes implementaciones de ADABOOST y SVM con una matriz como base de entrada, conformada por las características extraídas mediante texturas locales, *Local Binary Pattern* (LBP, en inglés).

En [3], estos métodos fueron comparados con imágenes de 3 tamaños diferentes: 24x24, 36x36 y 48x48 pixeles. La mejor tasa de clasificación fue alcanzada con el algoritmo SVM con imágenes de 36x36 pixeles, obteniendo una tasa de aciertos de 86.54 %, sobre la base de datos FERET. También en [15] se propuso un método para mejorar la clasificación de género, aumentando el conjunto de entrenamiento, con imágenes desalineadas intencionalmente. El mejor resultado global para la clasificación de género sobre la base de datos FERET fue de 92.5 %.

Una aproximación en la clasificación de género utilizando texturas, formas e intensidad de los pixeles, usando diferentes escalas fue propuesto en [6]. El mejor resultado en la clasificación de género basado en la intensidad de los pixeles fue de 87.85 %, para imágenes de tamaño 36x63 pixeles. Utilizando características de forma, la tasa de aciertos fue de 91.59 % de clasificaciones correctas con imágenes de tamaño de 128x128 pixeles y 93.46 % usando características de texturas LBP, también con imágenes de tamaño de 128x128 pixeles, todas utilizando la base de datos FERET.

Fusionando los 3 métodos utilizados (forma, intensidad y textura) se alcanzaron los mejores resultados para imágenes de 128x128 pixeles, con un 95.33 % de clasificaciones correctas en la base de datos FERET. En [6] se utilizaron varios tamaños de imágenes, como vectores de entrada de tamaño de 20x20, 36x36 y 128x128 pixeles. Cuando se fusionaron los 3 tipos de escalas y los 3 tipos de características se alcanzó una tasa de clasificación de 99.07 % con el conjunto de test, la cual fue obtenida en la base de datos FERET pero con un vector de entrada al clasificador de tamaño 1x18000 pixeles. Nuevamente el número total de entradas fue aumentado cerca de 9 veces con el uso de los 3 tipos de características y los 3 tipos de escalas.

El tiempo computacional aumenta directamente con el número de entradas al clasificador y es un factor importante de considerar para la aplicaciones en tiempo real que envuelven reconocimiento facial. Por lo tanto la selección de características es un proceso deseable.

Para problemas que tienen un número grande de características extraídas desde los datos de entrada, la selección de características es el proceso de seleccionar un subconjunto de características relevantes, que contenga información útil para distinguir una clase de las otras [16]. Uno de los principales objetivos de la selección de características es representar los datos en un espacio de baja dimensión [17]. La falta de un método efectivo para seleccionar características, ha sido compensado en parte mediante algoritmos de clasificación capaces de tratar al menos parcialmente las características con redundancia y relevancia [17–19].

El objetivo del proceso de la selección de características, es escoger el menor subconjunto de características que tienen la mayor cantidad de información sobre una posible clase. En [20] un método para la selección de características basado en la medición de la entropía para la clasificación de rostros fue propuesto utilizando el enfoque del algoritmo Local Gabor, el cual además de reducir significativamente el tiempo de cómputo, mejoró la tasa de reconocimiento

facial.

La información mutua ha sido usada como un criterio de selección de características, porque es una buena representación de la relevancia y redundancia entre variables aleatorias, considerando su robustez al ruido y a la transformación de datos [21].

La información mutua puede proveer un conjunto de características óptimas sin tener en cuenta los clasificadores [22]. La información mutua ha sido usada con éxito en otras aplicaciones de selección de características, del ámbito biomédico o de manejo de datos [22,23].

En esta tesis, se investigó la aplicación de los métodos de selección de características basados en información mutua (MI) para la clasificación de género y con el objetivo de reducir el número de características y mejorar las tasa de clasificación de género. Se compararon los resultados de 3 medidas de información mutua: Mínima redundancia y Máxima relevancia(mRMR), Información mutua normalizada (NMIFS) e Información Mutua Condicional (CMIFS). Se investigó también la fusíon de características de distinto tipo y su comportamiento al usar selección de características y se compararon los resultados con los métodos sin selección de características.

1.2. Objetivos

1.2.1. Objetivo general

El objetivo general de este trabajo es desarrollar un método para la clasificación de género, con bases de datos consolidadas internacionalmente que permitan incorporar algoritmos de selección de características, con la finalidad de poder determinar qué zonas del rostro son más importantes, al momento de clasificar.

1.2.2. Objetivos específicos

Los objetivos específicos son:

- Implementación de los métodos de clasificación de género de mejor desempeño previamente publicados. Evaluar resultados en bases de datos estándares, que permitan determinar el error en la clasificación de género.
- Aplicar los métodos de selección de características basados en información mutua a la clasificación de género. Comparar los resultados con y sin selección de características en bases de datos internacionales.

1.3. Organización de la tesis

En el Capítulo 2, se explican los conceptos teóricos necesarios para poder comprender la importancia de la selección de características, razones para disminuir la dimensión, los principales métodos de transformación lineal y una introducción a la teoría de la información. En el Capítulo 3, se estudian los métodos de selección de características mediante información mutua y sus distintas variantes. En el Capítulo 4, se explica la metodología de trabajo para la clasificación de género utilizada, con sus distintos clasificadores. Para finalmente en el Capítulo 5, analizar los resultados mediante tasas de clasificación para los diferentes métodos, como el tiempo empleado en cada tarea. Lo anterior con la finalidad de compararse con los métodos publicados y establecer conclusiones.

Capítulo 2

Selección de Características

Los métodos de selección de características pueden ser clasificados en 2 tipos: filtros y envolventes, los primeros son clasificadores sin conocimiento previo y no se dedican a un tipo específico de clasificación, ya que se basan en las propiedades intrínsecas de los datos, siendo independientes del proceso de aprendizaje. Por el contrario, las envolventes se basan en el desempeño de un tipo de clasificador, para evaluar la calidad del conjunto de características seleccionadas, por lo cual son dependientes del proceso de aprendizaje.

Las técnicas de aprendizaje estadístico habitualmente procesan conjuntos de N muestras o ejemplos de una variable D dimensional, que conforman una matriz de la forma $X = \{x_1, x_2, ...x_n\}, xi_i \in R^D$. Distintas maneras de distribución de datos y de su muestreo sugerirán el uso de una técnica para su análisis. Por ejemplo, se podría asumir, que hayan sido idénticamente distribuidos, dependiendo si las x_i pueden o no considerarse realizaciones de una variable aleatoria generada de acuerdo con la función densidad de probabilidad (FDP) $p(x)$, sea está conocida o no. Igualmente, se podrá suponer independencia o no entre las muestras, en función de si la probabilidad con que ha sido generada la muestra x_i depende de la muestra $x_j,\ j \neq 1$.

Las técnicas de aprendizaje de máquina presentes en la literatura suelen asumir que las muestras son independientes e idénticamente distribuidas, lo cual está justificado en muchas de sus aplicaciones.

Los datos contenidos en la matriz X pueden estar acompañados por datos adicionales dados por una segunda matriz $Y = \{y_1, y_2, ..., y_n\},\ y_i \in R^r$. En este caso, el problema de aprendizaje consistirá en inferir el valor de y a partir del de x. Si esta variable auxiliar es unidimensional y además adopta valores discretos, $y_i \in \{c_1, c_2...c_{nc}\}$, el problema de inferir su valor recibe el nombre de clasificación. Si la variable y es continua, el problema se denomina regresión. El caso general de y multidimensional se denomina multiregresión.

En cualquiera de los casos, el aprendizaje tiene lugar a partir de un conjunto de entrenamiento etiquetado, es decir, un conjunto de muestras para la cual conocemos los valores

de y. El conocimiento adquirido a partir de este conjunto muestral servirá para predecir el valor de la variable y en posteriores muestras de prueba no etiquetadas.

La investigación desarrollada en esta tesis se centra en el problema de clasificación. Con frecuencia, tanto el número de muestras N como el número de componentes (o dimensiones) D puede exigir una selección de características, tanto en el número de elementos como en sus componentes.

En el aprendizaje supervisado, la reducción se deberá realizar de forma que las muestras que se preservan son suficientemente representativas como para que podamos inferir propiedades extrapolables a las descartadas. También pueden eliminarse muestras que podamos asumir que no son representativas, porque corresponden a medidas erróneas de x.

La detección y eliminación de estas muestras atípicas (*outliers*, en inglés) ha sido estudiada para evitar el sesgo o deterioro en el proceso de aprendizaje. También se puede perseguir el objetivo de encontrar el subconjunto de muestras más representativo para trazar la función de clasificación o regresión. Algunos esquemas de aprendizaje, como los basados en SVM, realizan esa selección de datos implícitamente, ya que hacen uso de un criterio de máximo margen, que queda finalmente definido a partir de las muestras cercanas a la frontera de discriminación. Aun más, el problema de reducción de la dimensión de los datos ha recibido mucha más atención en las publicaciones [24].

2.1. Razones para reducir la dimensión

Los principales motivos para llevar a cabo una reducción en el número de variables o de un conjunto muestral son los siguientes:

- La capacidad de generalización de lo aprendido (es decir, la capacidad con que podremos aplicarlo a nuevas muestras) es mejor cuanto menor es la dimensión o cuando se compara con el número de muestras. Una alta dimensión conduce al sobre ajuste (*overfitting*, en inglés) [24], que constituye una adaptación excesiva del modelo de aprendizaje a los datos de entrenamiento, lo que impide que este sea difícil de extrapolar a nuevos datos.

- El costo computacional de la tarea de aprendizaje, el cual disminuye con la dimensión de los datos a los que se aplica. Estas motivaciones son comunes a los métodos de la selección de muestras como a los de extracción de características.

- El hecho de contar con una representación de los datos en baja dimensión facilita la interpretación de los mismos, así como su visualización.

2.2. Formas de reducción de la dimensión

La manera más sencilla de reducir la dimensión o número de componentes de los datos D es seleccionar un subconjunto de los mismos. En tal caso, se persigue que el nuevo conjunto de d componentes ($d \leq D$) conserve la información relevante sobre la estructura de los datos (en el caso del aprendizaje no supervisado) o sobre su capacidad para inferir la variable auxiliar y (en el caso del aprendizaje supervisado).

La selección de características busca el mejor subconjunto de variables de acuerdo con un determinado criterio. Algunos criterios exigen evaluar cada uno de posibles subconjuntos de variables para determinar el mejor de ellos. Este procedimiento es inviable por su complejidad, al no ser aplicable a datos que tengan un número de dimensiones apreciable, se suele acudir a un método secuencial para construir el mejor subconjunto.

Los métodos secuenciales más usados son la selección incremental (*Forward selection*, en inglés), la eliminación decremental (*backward elimination*, en inglés) y procedimientos basados en el algoritmo *branch-and-bound* [25]. En el caso de está tesis se trabajó bajo el criterio de búsqueda incremental.

Los métodos incrementales actúan de forma iterativa, añadiendo en cada paso una variable que se incorpora a las que ya han sido escogidas, proporcionando el valor máximo del criterio considerado. El esquema básico se resume a continuación.

Algoritmo 2.1 Algoritmo básico de la selección de características incremental

---

1. $S_0 = [\,]$;
2. $f_k = argMax_f C(f_k)$;
3. $S_1 = \{f_k\}$
   $Para\, n = 2\, hasta\, d$
4. $f_k = argMax_f C(S_{n-1} + f_k)$
5. $S_n = S_{n-1} + f_k$
   Fin

En el primer paso, se escoge la característica f_k que hace máximo el valor del criterio a $C(f_k)$. En problemas de clasificación, partimos de un conjunto vacio S_0, donde $C()$ viene dada por la precisión del clasificador o sobre algún conjunto de validación. A partir de ahí en cada iteración n, el conjunto óptimo de n características S_n estará formado por la unión del conjunto anterior S_{n-1} y la característica que unida a este conjunto, proporciona el mejor valor de $C(S_{n-1} + f_k)$.

La eliminación decremental opera de forma inversa: Se parte de los datos sin reducirlos e iterativamente se van eliminando variables. En cada caso se elimina la característica que

menos degrada el valor del subconjunto resultante. El procedimiento se ilustra a continuación.

En este caso, partimos de un conjunto S_0 formado por todas las variables. En cada paso n, se eliminan las características cuya ausencia tiene menor incidencia en el valor de $C(S_{n-1}-f_k)$.

Algoritmo 2.2 Algoritmo básico de selección de características decremental

1. $S_0 = [f_1, f_2, ..., f_D]$
2. $f_k = argMax_f C(S_0 - f_k)$
3. $S_1 = \{S_0 - f_k\}$
   $Para\ n = 2\ hasta\ d$
4. $f_k = argMax_{fk} C(S_{n-1} - f_k)$
5. $S_n = S_{n-1} - f_k$
   Fin

La selección incremental es la que mayor relevancia tiene en selección de características, ya que es el procedimiento sugerido en una gran parte de los métodos propuestos en la literatura. Esto se debe a que es en general, menos costoso en cuanto a tiempo computacional que la eliminación secuencial: si el número de componentes a seleccionar es menor que la mitad del número de componentes iníciales, la búsqueda incremental necesita menos iteraciones que la decremental.

2.3. Reducción de dimensión mediante transformaciones lineales

El objetivo de la reducción de la dimensión es extraer un conjunto de valores que sea de la mayor utilidad para la interpretación de los datos así como la identificación de la estructura o variedad conforme a la que están generados. Si esta variedad puede ser descrita en un espacio de dimensión menor a la original, el objeto de la reducción de la dimensión es proyectar los datos en este nuevo espacio.

Los tres tipos de transformaciones lineales más ampliamente utilizadas para reducir las dimensiones de los datos son: el análisis de componentes principales (PCA), el análisis discriminante lineal (LDA) y el análisis de componentes independientes (ICA). Cada una de estas técnicas persigue objetivos distintos:

• PCA: Busca una transformación lineal que proporcione la mejor representación de los datos (en el sentido de mínimo error cuadrático) con un vector de igual o menor dimensiones que el original, sin considerar la información de clases.

• LDA: Busca una transformación lineal, sobre un espacio de menos dimensiones que

el original, que proporcione la mejor separación entre las clases, en el sentido de mínimos cuadrados, además de considerar las información (etiquetas) de dichas clases.

• ICA: Busca una transformación lineal que proporcione atributos estadísticamente independientes. De esta forma se eliminan redundancias en los datos y se consigue expresar el vector de atributos de la forma más compacta posible.

Es importante notar que, a diferencia de LDA, ninguna de las técnicas de análisis de componentes tiene en cuenta la clase de los patrones para determinar la transformación a aplicar sobre los datos. La orientación es, por tanto, no supervisada y en principio no está garantizado que proporcionen ventajas en problemas de clasificación (aunque en general suelen funcionar bien).

2.3.1. PCA

El análisis de componentes principales o PCA es una herramienta de análisis de datos y una forma clásica de poder reducir la dimensionalidad de los datos. En tareas de reconocimiento de patrones es muy importante la selección de un adecuado conjunto de características [26]. Las características no deben estar correlacionadas entre sí, de manera que cada una aporte nueva información para el proceso de clasificación. PCA se utiliza normalmente para reducir el conjunto inicial de características. Este conjunto de características es elegido en forma arbitraria, es decir, no se utiliza información previa para escoger el mejor conjunto de características. Dado un espacio n-dimensional de entrada (una representación de los individuos como vectores de tamaño n), se desea encontrar m nuevas características, es decir, una representación de los individuos como vectores de características de m componentes, que permitan representar el espacio de entrada, donde $m < n$. Dada la condición de que los vectores de características a encontrar sean ortogonales entre sí, el problema equivale a proyectar un espacio de dimensión n en uno de dimensión m, lo cual conlleva una reducción de dimensión [27] [28].

PCA es un método para realizar esta reducción de dimensión que mantiene la información intrínseca de los datos de entrada y reduce al máximo el error generado desde el punto de vista del error cuadrático medio. La idea central del método PCA consiste en realizar un cambio de coordenadas de tal forma que los nuevos ejes, ortogonales entre sí, se orienten en aquellas direcciones donde los datos de entrada presentan mayor varianza. Las proyecciones de los datos de entrada en los ejes del nuevo sistema de coordenadas corresponden a los componentes principales.

El primer componente principal se elige a lo largo de la dirección con máxima varianza. Este proceso se repite con todos los componentes principales. Los vectores propios de la matriz de covarianza cruzada o de correlación cruzada corresponden a las direcciones de los componentes principales. De este modo la dirección del k-ésimo componente principal es el vector propio correspondiente al k-ésimo mayor valor propio de la matriz de covarianza. Un

paso previo al cálculo de la matriz de covarianza es calcular el promedio de los vectores de entrada y restar el vector promedio a cada vector de entrada.

Para el caso particular de la clasificación de rostros mediante PCA, primero se efectúa el cálculo de los rostros propios (eigenfaces, en ingles) Figura **??** siguiendo los pasos que se señalan a continuación:

Paso 1: Obtener imágenes de rostros I_1, $I_2, ...I_M$ para ser utilizadas como conjunto de entrenamiento. Estas imágenes deben estar centradas y deben tener el mismo tamaño (P pixeles de alto por Q pixeles de lado, con P y Q enteros)

Paso 2: Representar cada imagen I_i como un vector $\overrightarrow{\Gamma_i}$. En total se obtienen M vectores. Cada vector tiene $N = P \times Q$ elementos.

Paso 3: Calcular el vector promedio $\overrightarrow{\Psi}$:

$$\overrightarrow{\Psi} = \frac{1}{M} \sum_{i=1}^{M} \overrightarrow{\Gamma_i} \tag{2.1}$$

Paso 4: Restar al valor promedio:

$$\overrightarrow{\Phi} = \overrightarrow{\Gamma_i} - \overrightarrow{\Psi} \tag{2.2}$$

Paso 5: Calcular la matriz de covarianza C:

$$C = \frac{1}{M} \sum_{n=1}^{M} \overrightarrow{\Phi_n} \overrightarrow{\Phi_n^T} = A * A^T \tag{2.3}$$

donde $A = \left[\overrightarrow{\Phi_1}, \overrightarrow{\Phi_2}....\overrightarrow{\Phi_M}\right]$es una matriz de dimensión $N \times M$ y por tanto, C tiene una dimensión de $N \times N$.

Paso 6: Calcular los vectores propios $\overrightarrow{u_i}$ de la matriz $C = A * A^T$. Para imágenes típicas de rostros, calcular los valores propios y los vectores propios de la matriz C tiene un costo computacional alto. Por ejemplo, para imágenes de dimensión 100x200 pixeles se tendrían que calcular 20000 valores propios y 20000 vectores propios. Si la cantidad de individuos en la base de datos es menor a la cantidad de pixeles en una imagen ($M < N$), entonces existen sólo $M-1$ vectores propios significativos, pues el resto de los vectores propios están asociados a valores propios iguales a cero. Por lo tanto, basta con calcular los vectores propios de una matriz de dimensión $M \times M$ y asociar los vectores propios de está matriz a los vectores propios de C.

Si se calculan los vectores propios $\overrightarrow{v_i}$ de una matriz $L = A^T * A$ de dimensión $M \times M$ se tiene que:

$$L\overrightarrow{v_i} = A^T A \overrightarrow{v_i} = \mu_i \overrightarrow{v_i} \tag{2.4}$$

Si se pre-multiplican ambos lados de la ecuación por A se tiene que:

$$A * A^T = A * \overrightarrow{v_i} \tag{2.5}$$

lo que equivale a:

$$C\, A\, \overrightarrow{v_i} = \mu_i A \overrightarrow{v_i} \tag{2.6}$$

Por lo tanto, se tiene que $A\overrightarrow{v_i}$ son los vectores propios de la matriz C.

Se define entonces la matriz $L = A^T * A$ de dimensión $M \times M$, donde $L_{mn} = \overrightarrow{\Phi_m^T}\overrightarrow{\Phi_k}$ y se calculan los M vectores propios $\overrightarrow{v_i}$ de la matriz L. Los vectores propios $\overrightarrow{v_i}$ de la matriz L se relacionan linealmente con los vectores propios μ_i de la matriz C:

$$\overrightarrow{\mu_i} = \sum_{k=1}^{M} \overrightarrow{v_{ik}}\overrightarrow{\Phi_k},\ i = 1, ..., M \tag{2.7}$$

donde v_{ik} es el k-ésimo elemento del vector $\overrightarrow{v_i}$.

Paso 7: seleccionar los M' vectores propios asociados a los M' valores propios más altos ($M' < M$) para reducir la dimensionalidad. Los rostros propios (eigenfaces) calculados definen lo que se denomina un "espacio de rostros" que corresponde a un subespacio dentro del espacio de todas las imágenes. A modo de ejemplo, en la Figura**??** se muestra el conjunto de imágenes de entrenamiento utilizando los rostros propios (eigenfaces) calculados. Algunas aplicaciones concentran su atención en las proyecciones asociadas a los valores más pequeños aplicadas a rostros, que determinan el denominado subespacio de ruido, combinándolas con algoritmos genéticos para poder determinar el conjunto de caras propias que obtiene el mejor resultado de clasificación de género [29].

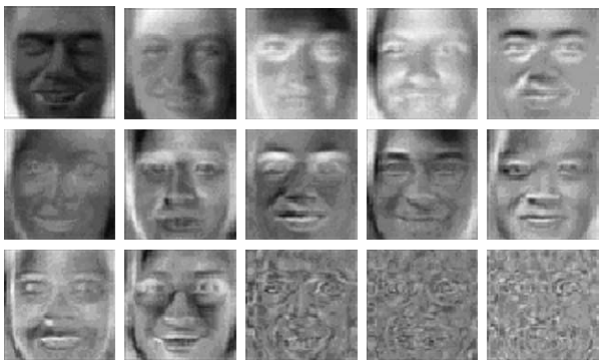

Figura 2.1: Ejemplos de rostros propios obtenidos con PCA, Fig. extraída de [1]

2.3.2. LDA

Dado un problema de clasificación multiclases con c clases y p ejemplos, se tiene que $\{X_i\}_{i=1}^{p}$, donde LDA [27] entrega una proyección lineal de los ejemplos iníciales sobre el subespacio de dimensiones d= c−1, maximizando la razón entre cada clase como su separación al interior de la clase. La base del subespacio transformado es $\{X_i\}_{i=1}^{d}$ el cual es obtenido maximizando la función de costo J.

$$J(w) = \sum_{i=1}^{d} \frac{w_i^T * S_B * w_i}{w_i^T * S_w * w_i} \tag{2.8}$$

donde S_B es la matriz de dispersión entre clases y S_w las matriz de dispersión de un ejemplo con los de su clase. El valor máximo está dado por los valores propios de $S_B w = S_w w * D$, donde w es la matriz cuyas columnas son w_i y D es la matriz diagonal de valores propios.

$$S_B = \sum_{c} N_c(\mu_c - \mu)(\mu_c - \mu)^T \tag{2.9}$$

$$S_W = \sum_{c} \sum_{i \in c} (x_i - \mu_c)(x_i - \mu_c)^T \tag{2.10}$$

siendo μ_c la media de cada clase, μ la media de todos los datos, N_c la cantidad de patrones de la clase c. LDA busca encontrar el vector w de proyección que maximice el "cociente" entre la matriz de dispersión inter-clase y la matriz de dispersión intra-clase. En consecuencia, para maximizar la solución debemos considerar el vector propio con mayor valor propio asociado.

En algunas aplicaciones, la naturaleza no negativa del problema hace que la descomposición de un vector en sus componentes principales no refleje la naturaleza no aditiva de los datos, por ejemplo, en tratamiento digital de imágenes, los valores de los pixeles son no negativos y la imagen puede ser expresada como una composición de los elementos que la conforman. La principal limitación de estos métodos es su incapacidad de hacer frente a situaciones en que los datos se articulan conforme a una estructura subyacente que sólo puede ser descrita mediante proyecciones no lineales.

2.3.3. ICA

El objetivo fundamental del Análisis de Componentes Independientes (ICA) [30], es el de proporcionar un método que permita encontrar una representación lineal de los datos no gaussianos, de forma que las componentes sean estadísticamente independientes o lo más independiente posible. Una representación de este tipo permite obtener la estructura fundamental de los datos en muchas aplicaciones, incluidas la extracción de características y la separación de señales.

Dado un conjunto de observaciones de variables aleatorias $\{x_1(t), x_2(t), ...x_n(t)\}$, siendo t el tiempo o el índice de las muestras, asumimos que están generadas por una combinación lineal de componentes independientes:

$$\begin{pmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{pmatrix} = \begin{pmatrix} s_1(t) \\ s_2(t) \\ \vdots \\ s_n(t) \end{pmatrix} \tag{2.11}$$

o en forma matricial:

$$x = A \cdot s \tag{2.12}$$

donde A es una matriz de mezcla desconocida. El Análisis de Componentes Independientes consistirá ahora en estimar tanto la matriz A como las fuentes $s_i(t)$ a partir de las observaciones $x_i(t)$. Supondremos que el número de observaciones coincide con el de las fuentes originales, si bien esta simplificación no es completamente necesaria para resolver el problema. De forma alternativa, podríamos definir ICA, como el problema de la obtención de la transformación lineal dada por la matriz W, tal que las variables aleatorias estimadas $y_i(t)$ con $i = 1, ..., n$ sean tan independientes como sea posible. Este planteamiento no difiere en exceso del original ya que una vez obtenida la matriz A, la matriz W se obtiene invirtiéndola.

El principio estadístico utilizado para determinar la matriz W es la independencia, es decir, las componentes y_i son estadísticamente independientes unas de otras, lo que significa que el valor que tome cualquiera de ellas no da información alguna sobre el valor que pueda tomar el resto. Esto resulta sencillo si los datos tiene una distribución Gaussiana, dado que es bastante simple encontrar componentes que sean independientes en este caso, atendiendo a que para datos Gaussianos, las componentes decorrelacionadas son siempre independientes [30]. Sin embargo, en la realidad los datos no suelen seguir una distribución Gaussiana, y la situación no es tan simple como estos métodos asumen. Muchos conjuntos de datos del mundo real tienen distribuciones supergaussianas, lo que significa que dichas variables aleatorias toman con mayor probabilidad valores que son cercanos al cero o valores muy grandes, en otras palabras, la función densidad de probabilidad ('pdf') de estos datos es puntiaguda en el cero y tiene las colas densas (debido a los valores grandes que toma lejos del cero), si la comparamos con la pdf de una variable Gaussiana de la misma varianza.

Para definir ICA de una forma rigurosa, es posible usar un modelo de variables ocultas [30]. Se trata de observar n variables aleatorias $x_1, ..., x_n$, que se modelan como una combinación lineal de las fuentes $s_1, ..., s_n$:

$$x_i = a_{i1} \cdot s_1 + a_{i2} + ... + a_{in} \cdot s_n \;\; \forall\, i = 1, ..., n \tag{2.13}$$

donde los a_{ij} son coeficientes reales. Por definición los s_i son independientes entre si. Este es el modelo básico ICA, que describe como las variables observadas son generadas por un proceso de mezcla de las fuentes s_j. Las componentes s_j son variables ocultas ya que no se pueden observar de forma directa. Además los a_{ij} pertenecientes a la matriz de mezcla se suponen también desconocidos. Las únicas variables que están 'visibles' serán las x_{ij} a partir de las cuales tendremos que estimar las fuentes s_j y la matriz de mezcla A. Este problema se tendrá que resolver de la manera más general posible.

2.4. Criterios basados en la teoría de la información

La teoría de la información de Shannon de mediados del siglo XX vino a cuantificar los conceptos de información o incertidumbre, así como describir de forma matemática la transmisión de información a través de un canal de comunicaciones. Aparte de las telecomunicaciones, ésta teoría ha tenido una profunda influencia en estadística y economía [31] .

En el campo de la reducción de la dimensión, la teoría de la información también ofrece un soporte indiscutible, dado que el problema a resolver en extracción y selección de características, es obtener características relevantes así como poco redundantes entre si. Esta teoría dispone de las herramientas necesarias para cuantificar el grado de dependencia entre cada característica y la variable de referencia (las clases) así como entre las propias características.

2.4.1. Introducción a la información mutua

La teoría de la información define el grado de dependencia estadística entre dos variables aleatorias a partir del concepto de información mutua. El nombre hace referencia al hecho de lo que se mide es la información que contiene una variable acerca del valor de alguna otra. La información mutua puede definirse a partir de la divergencia de Kullback-Leibler (KL) [31]. Esta divergencia viene dada, para dos distribuciones de probabilidades discretas P y Q con el mismo dominio X, $Y \in \{x_1, x_2, ...x_L\}$, por:

$$D_{K,L}\left(P,Q\right) = E_{P(X)}\left\{\frac{P(x)}{Q(x)}\right\} = \sum_{l=1}^{L} P(X_l) log \frac{P(x_l)}{Q(x_l)} \tag{2.14}$$

En donde, la información mutua entre las variables X e Y se define como la divergencia KL entre la distribución conjunta $P_{XY}(x,y)$ y el producto de las distribuciones marginales $P_X(x)$ y $P_Y(y)$:

$$I(X,Y) = D_{KL}(P_{XY}, P_X, P_Y) = \sum_{l_x=1}^{L_X} \sum_{l_y=1}^{L_Y} P_{XY}(x_{lx}, y_{ly}) log \frac{P_{XY}(x_{lx}, y_{ly})}{P_X(x_{lx}) P_Y(y_{ly})} \tag{2.15}$$

y que ahora X e Y pueden tener distinto dominio, $X \in \{x_1, x_2, ...x_{Lx}\}, Y \in \{y_1, y_2, ...x_{LY}\}$. La información mutua mide el grado de dependencia estadística entre 2 o más variables. Es por tanto una medida de reducción de incertidumbre que tenemos sobre el valor de una variable una vez que conocemos la otra.

La Entropía o autoinformación se define sobre una única variable cuando está es discreta, como:

$$H(X) = I(X,X) = -\sum_{l=1}^{L} P_X(x_l) log P_X(x_l) \tag{2.16}$$

La medida de Información Mutua mide la dependencia entre variables o conjuntos de variables. Si lo que estamos midiendo es la dependencia entre una variable y la clase, cuanto mayor sea el valor de la Información Mutua mayor será la dependencia existente entre las variables, ver la Figura **??**. Nosotros vamos a utilizar esta medida que podrá ser aplicada a problemas reales como es el caso de selección de características o bien la optimización de particiones. La incerteza de una variable C puede ser medida por la entropía $H(C)$. Para 2 variables aleatorias Y y C, la entropía condicional está definida por $H(Y/C)$, la cual mide la incerteza sobre la variable C dado que la variable Y es conocida.

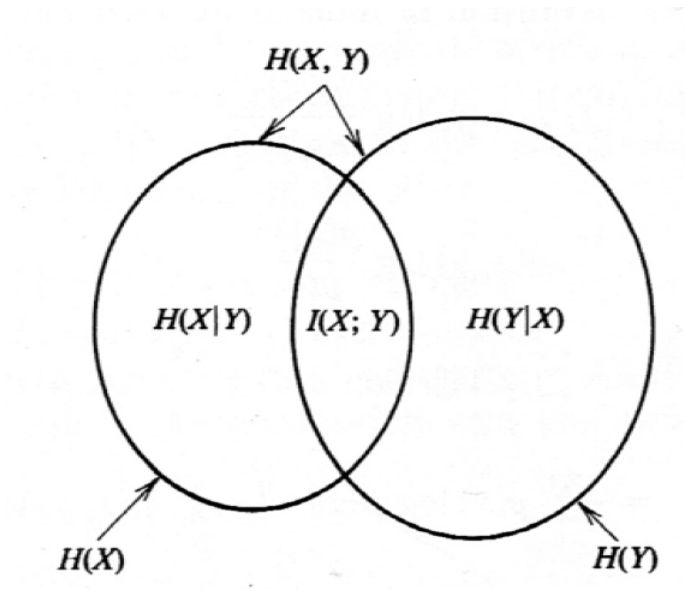

Figura 2.2: Relación entre Entropía e Información Mutua, Fig. extraída de [2]

Nuestro principal interés en este trabajo es diseñar un sistema de clasificación de forma

eficiente, considerando los enfoques de filtros y envolventes, teniendo en cuenta su estadística, tiempo computacional y tasa de clasificación. El filtro más usado en la actualidad que ordena las características de acuerdo su importancia, es la información mutua. Sus propiedades más importantes son su habilidad para cuantificar la dependencia no lineal entre características y su invariancia ante la transformación de espacios de datos.

La selección basada en un ordenamiento (ranking) nos asegura la dependencia débil entre las características y puede llevar a que las características seleccionadas sean redundantes y por lo tanto aporten menos información.

El cálculo o estimación de la información IM, debe considerar el cálculo de la estimación de la funciones de probabilidad $p(x_i), p(y_i)$ y $p(x_i, y_i)$ las cuales son desconococidas a priori y deben ser estimadas de los mismos datos. Para poder superar este problema, los valores se discretizan o aproximan sus densidades con métodos paramétricos o no paramétricos. Este problema presenta una solución muy práctica ya que se trabaja con sumatorias, donde las probabilidades son estimadas desde la cuenta de frecuencia de los datos o histogramas. Las publicaciones actuales han tratado de solucionar este problema desde diferentes ámbitos con la finalidad de no tener que estimar demasiados parámetros [18,32].

2.4.2. Estimación de la función de probabilidad (FDP)

De acuerdo a lo indicado por la teoría de la información [31], se puede decir que se necesita disponer de la FDP de las variables con las que trabaja. En un ámbito de aprendizaje basado en muestras, los datos se interpretan como realizaciones de una variable aleatoria con una densidad desconocida. En aprendizaje, se asume la diferenciación entre los métodos generativos y discriminativos. Los primeros hacen uso de los modelos de función de probabilidades de los datos, mientras que los últimos no necesitan dichos datos. La estimación de la FDP de los datos es necesaria si se pretende hacer uso de la información mutua o alguna otra de las divergencias explicadas, como criterio para reducir dimensión.

Los métodos propuestos en la literatura se suelen dividir en tres grupos: paramétricos, semi-paramétricos y no paramétricos. Los métodos paramétricos son los más sencillos de usar y asumen que la densidad desconocida pertenece a una determinada familia de FDPs con descripción analítica. La tarea de estimar la FDP se reduce, por tanto, a encontrar los parámetros del modelo que mejor se ajustan a los datos. Este ajuste suele llevarse acabo mediante un criterio de máxima verosimilitud [24].

Los modelos semi-paramétricos no se restringen a una forma funcional en concreto, por lo que son más flexibles. Los denominados "Modelos de mezcla" responden a esta descripción y consisten en mezclas de modelos paramétricos que tienen la forma:

$$p(x) = \sum_{l=1}^{L} \alpha_l f(x \mid \theta_l) \tag{2.17}$$

donde cada α_l es la probabilidad a priori de que x pertenece al grupo l. Se cumple que $\sum_l \alpha_l = 1$. Cada submodelo está caracterizado por un conjunto de parámetros θ_l.

En concreto, los modelos basados en Mezclas de gaussianas (*Gaussian Mixture Models*, GMM en inglés) [33] por ejemplo, constituyen la familia más extendida de modelos de mezcla. Un modelo GMM asume un agrupamiento de datos, de forma que cada muestra pertenece a un grupo que responde a un modelo de tipo paramétrico gausiano. El problema de estimar simultáneamente los α_l y los θ_l es resuelto satisfactoriamente por el algoritmo de Maximización de la Esperanza (*Expectation Maximization*, EM en inglés) [24].

En el aprendizaje basado en teoría de la información, los GMM tiene el problema de que si cambiamos el conjunto de variables a utilizar, en un esquema secuencial de selección o extracción de características, se exige reajustar el modelo en cada iteración, que es algo que puede llegar a ser computacionalmente muy costoso. Además persiste, aunque en menor grado, la dependencia de la validez del modelo con la familia de modelos paramétricos considerados para la mezcla.

Estas limitaciones han hecho que los modelos no paramétricos sean más utilizados en aprendizaje basado en teoría de la información. El esquema no paramétrico más difundido son las ventanas de Parzen [34]. Este tipo de modelo recibe también el nombre de densidad basada en núcleos (*Kernel density estimation*, KDE en inglés).

El modelo de Parzen tiene la siguiente forma:

$$p(x) = \frac{1}{N} \sum_{i=1}^{N} k(x - x_i) \tag{2.18}$$

donde k es la ventana o Kernel utilizada. La función ventana tiene las mismas propiedades que una FDP, es decir, es no negativa y tiene una integral unitaria a lo largo de su dominio. El modelo se construye a partir de un conjunto de muestras o centroides x_i, que son realizaciones de la variable aleatoria cuya FDP se pretende modelar. En estos puntos es donde están centradas las ventanas.

Lo que hace que este modelo sea tan ampliamente usado es el hecho de que dependa únicamente de muestras y de la elección de la función Kernel. En realidad, a pesar de su denominación de método no paramétrico, en un modelo de Parzen hay que precisar el tipo de ventana a usar y su ancho. Un inconveniente de estos métodos es el alto costo de evaluar la probabilidad en un punto x determinado, ya que exige la evaluación de N ventanas, lo que puede ser intensivo para base de datos con gran tamaño de muestras o ejemplos.

Capítulo 3

Métodos de Selección de Características con Información Mutua

La Información Mutua I [32] entre 2 variables x e y se define en base a su probabilidad de distribución conjunta $p(x, y)$ y sus respectivas probabilidades marginales $p(x)$ y $p(y)$ como:

$$I(x, y) = \sum_{i,j} p(x_i, y_j)\, log \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \tag{3.1}$$

Si un pixel tiene una expresión aleatoria o distribuida uniformemente en diferentes clases, su Información Mutua con esas clases es cero. Por el contrario, si un pixel se diferencia fuertemente y es expresado en diferentes clases, se puede decir que su Información Mutua es grande, de tal forma se puede indicar que se usa Información Mutua para medir el nivel de similitud entre pixeles. El concepto de mínima redundancia($min\, W_I$) , como en la eq (3.2), permite seleccionar pares de pixeles que son muy diferentes en términos de Información Mutua [32]

$$min\, W_I, WI = \frac{1}{\mid S \mid^2} \sum_{fi,fs \epsilon S} I(fi, fs) \tag{3.2}$$

donde S muestra el subconjunto de característica y $\left|\mathrm{S}\right|$ es el número de características en S, e $I(fi, fs)$ es usado para representar la información mutua entre $I(pixel(i), pixel(s))$. La capacidad discriminadora para las diferentes clases puede ser obtenida por la Información Mutua entre las clases $c = \{c_1, c_2, c_3....c_k\}$ y el $pixel(i_i)$ dado por $I(c, i_i)$, así $I(c, fi)$ mide la relevancia del fi_{esimo} pixel para la tarea de clasificación. Por lo tanto, la condición de máxima relevancia($max\, V_I$), maximiza la relevancia total de todos los pixeles en S [32] como:

$$max\, V_I, V_I = \frac{1}{\left|S\right|} \sum_{fi \epsilon S} I(c, fi) \tag{3.3}$$

Las primeras características son seleccionadas de acuerdo a V_I, por ejemplo la característica con la más alta $I(c, fi)$, las demás características son seleccionadas incrementalmente desde el conjunto de características S. Si m características ya han sido seleccionadas desde S y una característica adicional es seleccionada desde $\Omega_s = \Omega - S$, luego las 2 condiciones son optimizadas, la operación Min es interpretada como mínima redundancia y la operación Max es interpretada como máxima relevancia.

3.1. Mínima redundancia máxima relevancia (mRMR)

Tradicionalmente, la teoría de la información es usada para cuantificar los conceptos de Relevancia y Redundancia:

3.1.1. Relevancia

Teniendo como entrada un conjunto de características F y clases de salida c, hay que encontrar cuales características tienen más información que permita describir las clases c, pero la decisión de cuales características deben ser escogidas está asociada al grado de dependencia de cada características individualmente, pero puede ocurrir que un grupo de características sea más importante que una sola característica actuando aisladamente, esto implica que hay niveles de relevancia entre ellas.

3.1.2. Redundancia

El termino redundancia está asociado a la magnitud de la dependencia entre 2 o más características en F, la cual puede ser cuantificada por la información común compartida entre características.

La redundancia tiene las propiedades de ser: no lineal, simétrica, no negativa y no decreciente con el número de de características [23].

Hay 2 formas de poder combinar estos conceptos operaciones de relevancia y redundancia [23,32], Información Mutua diferencial MID e Información Mutua de cociente MIQ como:

$$MID = max(V_I - W_I) \tag{3.4}$$

$$MIQ = max(V_I/W_I) \tag{3.5}$$

El conjunto de características de mRMR es obtenido optimizando la condición en (3.4)y (3.5) simultáneamente. La optimización de ambas condiciones requiere combinarlas en una sola función [32] como:

$$f^{mRMR}(X_i) = I(c; fi) - \frac{1}{\mid S \mid} \sum_{fi \epsilon S} I(fi; fs) \tag{3.6}$$

donde , $I(c; fi)$ mide la relevancia de la característica agregada para la clase c y el término $\frac{1}{|S|} \sum_{fi \epsilon S} I(fi; fs)$ estima la redundancia del $fi_{ésima}$ característica respecto al conjunto de características previamente seleccionadas S.

3.2. Información mutua normalizada (NMIFS)

En [35], fue propuesta una mejora al algoritmo mRMR basado en Información Mutua normalizada de las características, limitando la Información Mutua de 2 variables a través del mínimo de sus entropías. Como la entropía de una característica puede variar enormemente, su medida debe estar normalizada antes de aplicarlas a un conjunto global de características [35] como:

$$f^{NMIFS}(X_i) = I(c; fi) - \frac{1}{\left|S\right|} \sum_{fi \epsilon S} I_N(fi; fs) \tag{3.7}$$

donde $I_N(fi; fs)$ es la Información Mutua normalizada por el mínimo de las entropías de ambas características , la cual es definida por:

$$I_N(fi; fs) = \frac{I(fi; fs)}{min(H(fi), H(fs))} \tag{3.8}$$

Luego c es la clase de salida y S conjunto de características seleccionadas, Ω es el conjunto de características candidatas y $fi \, \epsilon \, \Omega$.

Donde fi es la $fi_{ésima}$ característica seleccionada.

Algoritmo 3.1 Algoritmo de selección de características mediante NMIFS.

1. Se inicializa: el conjunto $\Omega = \{fi/i = 1, ...(N)\}$, con N características y $S = \{\emptyset\}$, conjunto vacio.

2. Se calcula la I con respecto a cada clase; calculando $I(fi; c)$ para cada característica $fi\,\epsilon\,\Omega$.

3. Se selecciona la primera característica, con $fi = max_{i=1,...N}\{I(fi; c)\}$ desde el conjunto $\Omega_s \leftarrow \Omega/\{fi\}$;con $S \leftarrow \{fi\}$.

4. Selección algoritmo de tipo greedy, se repite hasta $\mid S \mid = k$, donde k es el número de características solicitadas, luego:

   a) Se calcula la I entre características; $I(fi; fs)$ para todos los pares $fi; fs$ con $fi\,\epsilon\,\Omega$ y $fs\,\epsilon\,S$, si no está disponible

   b) Selecciona la próxima característica $fi\,\epsilon\,\Omega$ que maximiza la medida (3.7), desde que $\Omega_s \leftarrow \Omega/\{fi\}$y $S \leftarrow \{fi\}$.

5. Se obtiene el conjunto S que contiene las características solicitadas.

3.3. Información mutua condicional (CMIFS)

Sea S el conjunto de características ya seleccionadas y Ω el conjunto de características candidatas, $S \cap \Omega = \emptyset$, y c las clases de salida, la próxima característica seleccionada de Ω será aquella que máxime $I(c; fi, S)$ donde $fi\,\epsilon\,\Omega$ y:

$$I(c; fi, S) = I(c; fi) - [I(fi; S) - I(fi; S \mid c)] \tag{3.9}$$

Luego c es la clase de salida y S conjunto de características seleccionadas, Ω es el conjunto de características candidatas y $fi\,\epsilon\,\Omega$.

Donde fi es la $fi_{ésima}$ característica seleccionada.

Algoritmo 3.2 Algoritmo de selección de características mediante CMIFS.

1. Se inicializa: el conjunto $\Omega = \{fi/i = 1, ...(N)\}$, con N características y $S = \{\emptyset\}$, conjunto vacío.

2. Se calcula la I con respecto a cada clase; calculando $I(fi; c)$ para cada característica $fi\,\epsilon\,\Omega$.

3. Se selecciona la primera característica que máxima la I, con $fi = max_{i=1,...N}\{I(fi; c)\}$ desde el conjunto $\Omega_s \leftarrow \Omega/\{fi\}$;con $S \leftarrow \{fi\}$.

   a) Selecciona la característica $fi\,\epsilon\,\Omega$ que maximiza la medida (3.9) hasta $\Omega == Null$

4. Desde que $\Omega_s \leftarrow \Omega/\{fi\}$ y $S \leftarrow \{fi\}$.

5. Se obtiene el conjunto S que contiene las características solicitadas.

Capítulo 4

Metodología

La clasificación de género tiene por finalidad separar la información de las imágenes de acuerdo a su clase, ya sea hombre o mujer. Esto se puede realizar de diferentes maneras, entregando al clasificador la información de la imagen como vector, sin ningún tipo de procesamiento o realizando preprocesamiento de imágenes, es decir, normalizando su iluminación, alineando las imágenes o extrayendo características con la finalidad de reducir el tiempo de procesamiento o como el resultado de una combinación lineal de datos que permitan representar la información de clase en otro espacio donde sea más fácil su separación.

Como se aprecia en el diagrama del proceso de clasificación en la Figura **??** , nuestro trabajo se realizó en 5 etapas principales. En un comienzo se buscó el rostro de una persona (hombre o mujer) en toda la imagen hasta ser detectada, para luego ser recortada y alineada de acuerdo a la coordenadas de sus ojos, posteriormente se redimensionó a 3 tamaños diferentes, para en su etapa final proceder a extraer las características más relevantes, mediante Información Mutua y ser clasificadas.

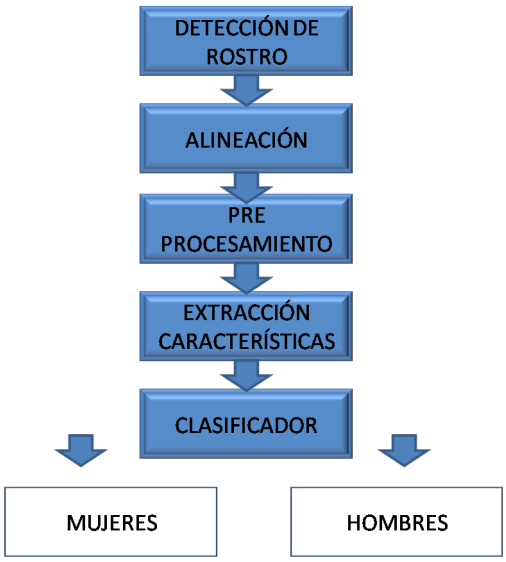

Figura 4.1: Diagrama del proceso de clasificación de género.

4.1. Detección de rostro y alineación

Para la etapa de detección de rostros se utilizó el software Face Detect 1.0 implementado en las librerías de OpenCv 2005, basado en características Haar, para formar un detector en cascadas [36]. Este detector busca en toda la imagen el rostro de una persona, como se aprecia en la Figura **??**, la menor imagen que puede ser detectada es de 24x24 pixeles. La salida del detector de rostros es la entrada a nuestro sistema de clasificación.

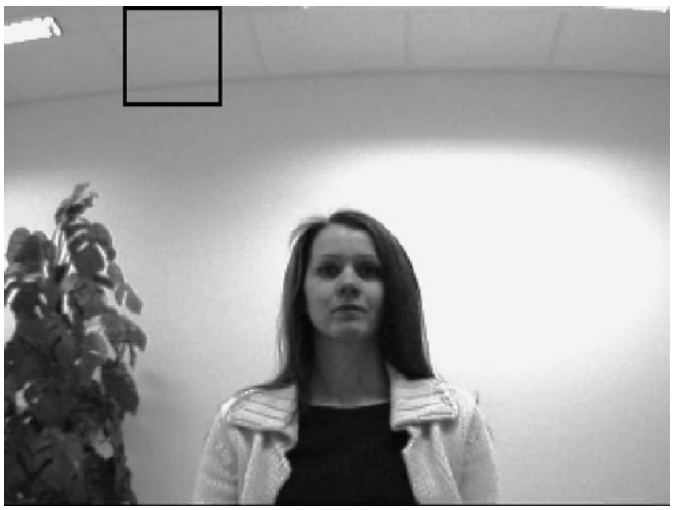

Figura 4.2: Ventana ADABOOST, buscando un rostro en la imagen.

En el caso de la alineación, los rostros fueron recortados de acuerdo a las siguientes proporciones faciales [3], lo cual se precia en la Figura **??**.

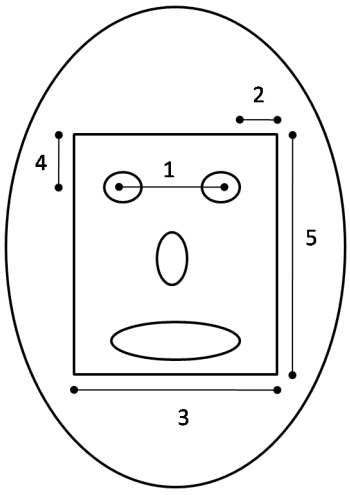

Figura 4.3: Medidas utilizadas para alinear y recortar las imágenes.

- El ancho del rectángulo recortado de la imagen (1) se determinó usando la distancia entre ojos d_e

- La distancia a los costados de los ojos (2) es: $d_e * 0{,}25$
- El ancho total del rectángulo (3) es: $d_e * 1{,}5$
- La distancia sobre los ojos (4) es: $d_e * 0{,}5$
- La altura del rectángulo (5) es: $d_e * 2{,}2$

Una vez detectados los rostros mediante características Haar, las cuales recorren la imagen en diferentes posiciones y escalas como se aprecia en la Figura 4.4, y alineados, estos fueron transformados en una matriz de $n \times m$, donde n es la imagen como vector y m el número de imágenes de la base de datos. Esta matriz es la entrada al método de selección de características, para luego ser procesado por el clasificador correspondiente. Aquellas imágenes que no pudieron ser detectadas se marcaron a mano.

4.2. Preprocesamiento

Se debe considerar que las imágenes de la base de datos FERET, son de muy buena calidad, tiene rostros frontales y con iluminación controlada relativamente homogénea, además posee las coordenadas las posición de los ojos, las cuales fueron utilizadas para rotar los rostros y dejar los ojos de todas las imágenes en la misma posición, solo se debió agregar la información de género, asignado un 1 para las mujeres y un 0 a los hombres.

Por el contrario, la base de datos WEB, tiene imágenes seleccionadas desde internet en los más diversos tamaños y calidades. No poseen las coordenadas de los ojos, por lo cual no se pueden alinear, se considera una base de datos bastante difícil y poco homogénea en cuanto a iluminación razón por la cual, se ecualizó su histograma antes de ser utilizada en los clasificadores.

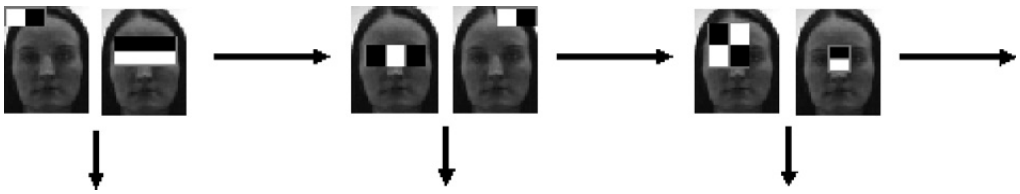

Figura 4.4: Ejemplo del proceso de detección de rostros mediante características Haar.

4.2.1. Ecualización de histograma

Los histogramas constituyen la base de varias técnicas de procesamiento de los tonos de grises [37]. La manipulación de los histogramas es usada de manera eficiente en el realce o mejoramiento de la calidad de una imagen. La información estadística obtenida a partir de los histogramas se utiliza en diversas aplicaciones como compresión y segmentación de

imágenes. La facilidad con la que se pueden calcular los histogramas usando software y su bajo consumo de recursos de hardware en su implementación, han hecho de esta herramienta una de las más usadas en el procesamiento en tiempo real.

El histograma de una imagen es la representación gráfica de la distribución que existe de las distintas tonalidades de grises con relación al número de pixeles o porcentaje de los mismos, es decir, un histograma representa la frecuencia relativa de ocurrencia de los niveles de gris. La representación de un histograma ideal sería la de una recta horizontal, ya que eso nos indicaría que todos los posibles valores de grises están distribuidos de manera uniforme en nuestra imagen.

La ecualización del histograma es una técnica bastante conocida y sirve para obtener un histograma uniforme de tal manera que los niveles de gris son distribuidos sobre la escala y un número igual de pixeles sean colocados en cada nivel de gris. Para un observador, esta ecualización hace que las imágenes se vean más balanceadas y con mejor contraste. Como consecuencia, una imagen ecualizada, permite que ciertos detalles sean visibles en regiones oscuras o brillantes, como se aprecia en la Figura 4.5.

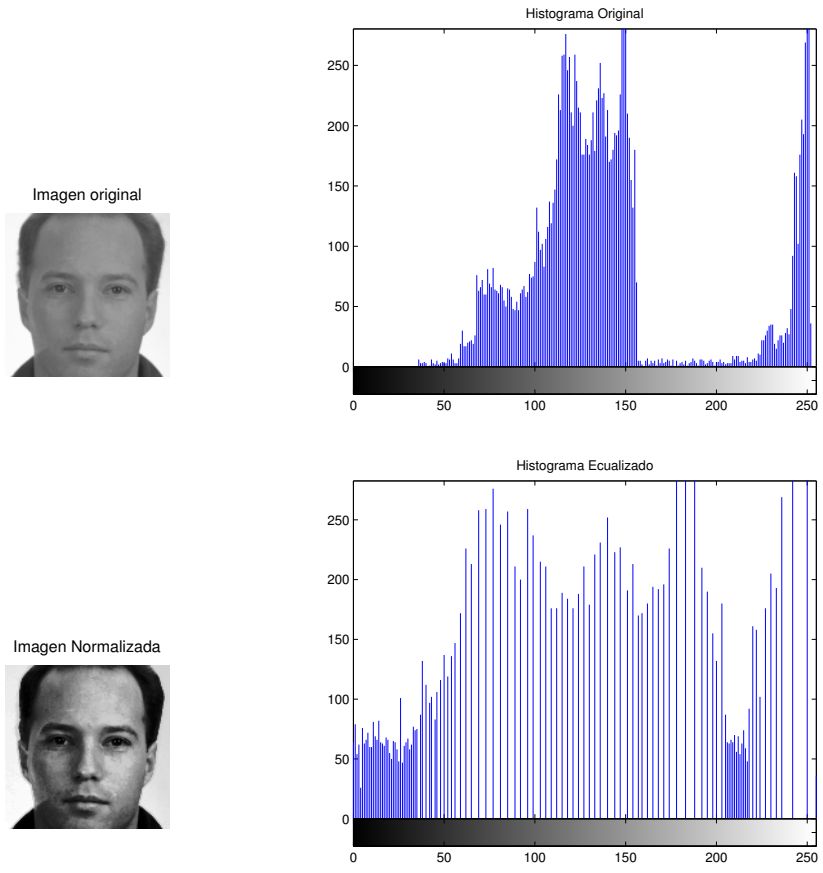

Figura 4.5: Arriba: Imagen original y su histograma, Abajo: Imagen normalizada y su histograma.

4.3. Extracción de características

La extracción de características se realizó mediante el uso de Información Mutua, aplicada a pares de pixeles en las imágenes recortadas luego de la etapa de detección de rostro, para luego ser alineadas y redimensionadas. Estas imágenes fueron transformadas en una matriz de $n \times m$, donde n es la imagen como vector y m el número de imágenes de la base de datos. Esta matriz es la entrada al método de selección de características. Se realizaron selecciones con 3 métodos: mRMR, NMIFS y CMIFS.

En el caso de la selección de las características de textura (LBP), se realizó el mismo procedimiento aplicado en [3]. La imagen del rostro, proveniente de la etapa de detección y redimensionamiento fue dividida en N sub-bloques de 8×8 pixeles y el operador LBP fue aplicado a cada sub-bloque usando $LBP_{4,1}$, es decir los 4 conectores vecinos y de radio 1 para luego aplicarlo a la imagen completa, mediante la configuración $LBP_{8,1}$. Ambos histogramas fueron concatenados para obtener el vector de características, (Figura 4.6), desde donde se seleccionaron las características con los métodos mRMR, NMIFS y CMIFS. Finalmente se formaron vectores de características relevantes, los cuales fueron las entradas a los diferentes clasificadores que se detallan en 4.4

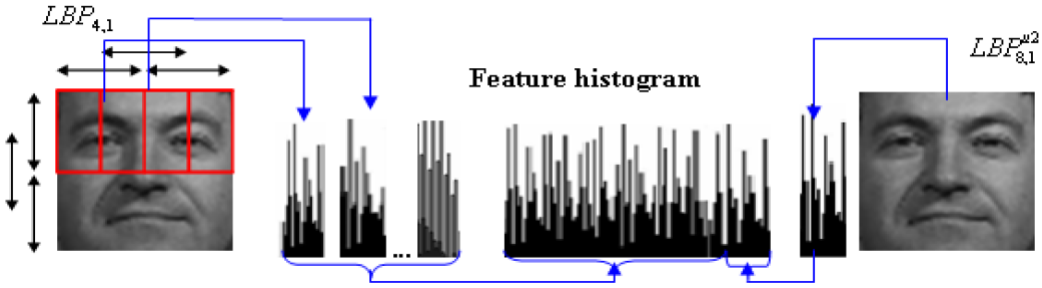

Figura 4.6: Representación gráfica de la aplicación de algoritmo LBP en imágenes faciales, Fig. extraída de [3].

4.4. Clasificadores

4.4.1. Redes neuronales

Una red neuronal es un arreglo de neuronas artificiales. Existen en la actualidad varios tipos de redes neuronales, siendo el más común el perceptron multicapa, en el cual las neuronas se distribuyen en capas, donde las entradas a cada capa son la salida de la anterior o señales externas. En la Figura 4.7 se puede apreciar un modelo de la estructura de una red neuronal del tipo perceptron multicapa, la cual se utilizó en este trabajo. Las entradas x_i son ponderadas con ciertos pesos w_{ij} al ingresar a cada neurona de la capa oculta, donde la suma de esta, es el argumento de una función de activación, que para este trabajo fue del

tipo tangente hiperbólica. Las salidas de cada neurona de la capa oculta son ponderadas por un peso determinado antes de ser sumadas y ser argumento de las funciones de activación de las neuronas de la capa de salida, la cual no es necesariamente la misma función de activación de las neuronas capas ocultas.

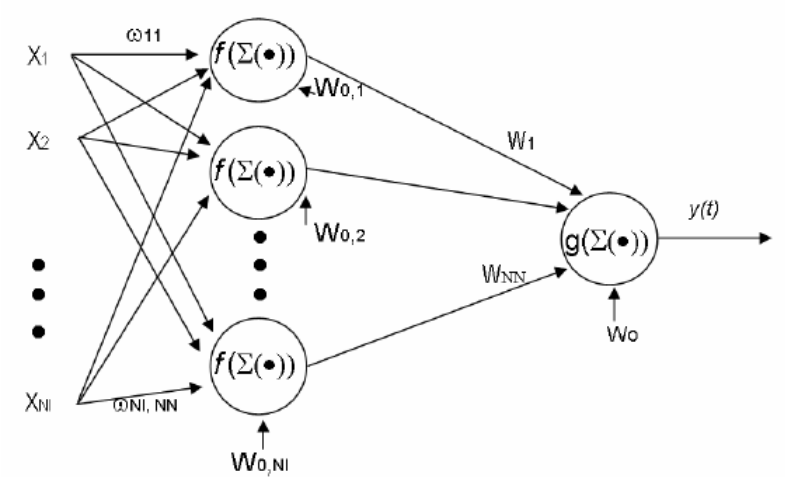

Figura 4.7: Esquema de una red neuronal de tipo perceptrón multicapa.

El modelo neuronal puede ser planteado por medio de una expresión matemática que relaciona los pesos sinápticos y las entradas para obtener la salida del proceso. La expresión matemática para la k-ésima salida de la red neuronal se aprecia a continuación:

$$y_k = g\left[\sum_{i=1}^{NN}\left(W_i.f\left(\sum_{j=1}^{NI} w_{ji}.x_j + w_{0i}\right)\right) + w_{0k}\right] \tag{4.1}$$

donde:

y: Salida del sistema

$g(.)$: Función de activación de las neuronas de la capa de salida

NN: Número de neuronas en la capa oculta

W_i: Peso desde la neurona i de la capa de oculta a la salida k

$f(.)$: Función de activación de las neuronas de la capa oculta

NI: Número de entradas al sistema

w_{ji}: Peso desde la neurona j de la capa de entrada a la neurona i de la capa oculta

x_j: Entrada j al sistema

w_{0i}: Sesgo de la neurona i de la capa oculta

w_0: Sesgo de la salida

Para poder utilizar las imágenes como entrada a las redes neuronales, se realizó un preprocesamiento a las imágenes de manera de ecualizar su histograma y compensar la intensidad de la iluminación de las imágenes, para luego crear un vector de características. Una red multicapa con imágenes ecualizadas fue usada, la intensidad de sus pixeles fue normalizada entre

-0.5 a 0.5 y almacenadas en un vector, la capa de entrada es igual al número de pixeles. Se probaron redes con 20,10,5,2 y 1 capa oculta, con función de activación tangente hiperbólica. Como salida se utilizó una capa de 1 neurona que evalúa valores entre -0.5 y 0,5. Todos los valores iguales o menores que -0.5 se clasifican como mujeres, los valores sobre 0.5 se evalúan como hombres.

4.4.2. Clasificador SVM(*Support Vector Machine*)

La Máquina de Soporte Vectorial o SVM fue desarrollada en 1995 por Vapnik y Lerneren en los laboratorios AT&T. Fue ideada originalmente para la resolución de problemas de clasificación binarios en los que las clases eran linealmente separables, ya que la solución obtenida era aquella en la que se clasificaban de manera correcta todas las muestras, colocando el hiperplano de separación lo más lejos posible de todas ellas. Para lograr esto, se realiza un entrenamiento con los datos dispuestos para ese fin, a partir del cual se define el hiperplano óptimo, el que maximiza el margen entre las muestras de cada clase respecto a la frontera de separación.

El objetivo es que luego del entrenamiento, la máquina generalice bien para datos nuevos que no han participado en el entrenamiento, es decir, clasifique correctamente a qué clase pertenecen cada una de las muestras. Para ello, intenta buscar una función que minimice el costo sobre el conjunto de entrenamiento, manteniendo a su vez la correcta generalización de la máquina.

Para obtener una buena generalización, hay que tratar de evitar dos de los efectos típicos del entrenamiento:

- Sobre-ajuste: los parámetros obtenidos en el entrenamiento se ajustan demasiado a las muestras por lo que se pierde generalización.
- Sub-ajuste: el hiperplano trazado es demasiado sencillo y no logra una buena generalización debido a que se han utilizado muy pocas muestras de entrenamiento, por lo que la máquina no generará buenos resultados con muestras nuevas.

Matemáticamente dado N muestras x_i, con sus respectivas etiquetas, que en nuestro caso indicarán si la imagen es un hombre o una mujer, y_i definirá a que clase pertenece:

$$(x_i \cdot y), (x_2 \cdot y_2), ..., (x_N \cdot y_N); x_i \in R^j, y_i \in \{+1, -1\} \tag{4.2}$$

siendo j el número de dimensiones o componentes de los vectores que contienen los datos.

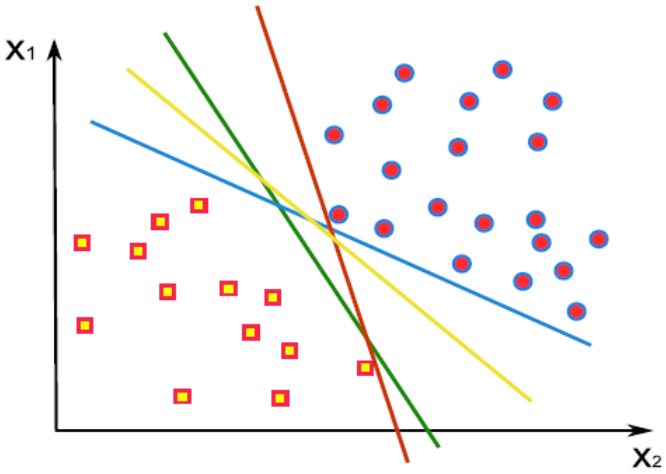

Figura 4.8: Clasificador convencional, linealmente separable.

4.4.2.1. Caso linealmente separable

Se dice que un problema es linealmente separable cuando para cualquier conjunto de muestras existe al menos un único hiperplano que clasifica los patrones con error cero. De todo, el conjunto de muestras, se extraen una serie de vectores, los vectores soporte, que son los únicos que se necesitan de entre todos los datos para definir la frontera de decisión, como se aprecia en la Figura 4.8

Un clasificador es lineal si su función de decisión puede expresarse mediante una función lineal en x, Así la ecuación del hiperplano de separación será el lugar de los puntos en x en los que se cumple:

$$H : x \cdot w + b = 0 \tag{4.3}$$

siendo b una constante que indica la posición del plano respecto al origen de coordenadas. Esta constante recibe el nombre del sesgo. Por otro lado$w$ es el vector normal al hiperplano.

4.4.2.2. Caso linealmente no separable

En el caso que el problema no sea linealmente separable, hay dos soluciones: buscar una frontera no lineal o tratar de encontrar el hiperplano que cometa un menor número de errores, ver la Figura 4.9. Para eso se introduce las variables positivas ξ que controlan el error permitido y penalizan las muestras mal clasificadas, así se puede indicar que:

$$y_i \{x_i \cdot w + b\} \geqslant 1 - \xi_i; \xi_i \geq 0, \forall i \tag{4.4}$$

así para las muestras bien clasificadas, se cumplirá que $0 < \xi_i < 1$, dependiendo de qué tan cerca este de la frontera la muestra, mientras que en las mal clasificadas ξ_i>1.

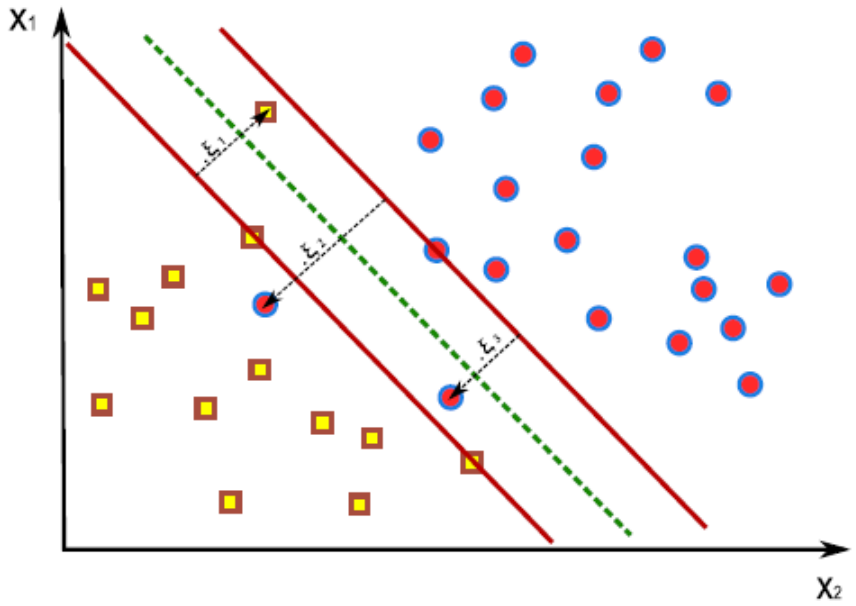

Figura 4.9: Clasificador con hiperplano de margen blando.

El algoritmo SVM fue aplicado de la misma manera que en [3], para clasificar género, es decir, se creó un vector de características, el cual en una primera etapa se utilizaron las imágenes completas, para luego compararlas con un vector creado mediante uno de los 3 vectores de características, seleccionando incrementalmente el número de ellas con la finalidad de probar su desempeño.

4.4.2.3. SVM no lineal

El clasificador SVM no lineal, se puede interpretar como una generalización del hiperplano óptimo de decisión, ya que permite la resolución de problemas no separables y trazar fronteras de clasificación no lineales, Figura 4.10 si para el caso en que los datos NO son linealmente separables, como en el caso de la clasificación de género, puede aplicarse una transformación $\phi(x)$ sobre el espacio de trabajo con el objetivo de obtener un espacio de características, generalmente de dimensión superior, donde si sean separables las muestras y donde se debe trazar el hiperplano óptimo de separación. La formulación es básicamente la misma, únicamente desplazamos x por $\phi(x)$. Para construir un SVM en un espacio resultante este debe cumplir:

$$k(x, x_i) =< \phi(x), \phi(x_i) > \tag{4.5}$$

donde k es llamada la función Kernel. Teniendo está función, es posible aplicar el algoritmo de entrenamiento del SVM sin conocer ϕ. Existen distintas funciones Kernel que permiten

adaptar el SVM a cada conjunto de muestras, con el fin de obtener mejores resultados, los más usados son:

- Lineal:

$$k(x, x_i) = x \cdot x_i \tag{4.6}$$

- Polinomial

$$k(x, x_i) = (\gamma x \cdot x_i + c)^{\alpha} \tag{4.7}$$

- Gausiano

$$k(x, x_i) = exp(-\gamma \mid x - x_i \mid^2) \tag{4.8}$$

  siendo γ una constante de proporcionalidad, c un coeficiente y α el valor del polinomio, en este trabajo de investigación se ha optado por un Kernel Gaussiano de acuerdo a lo indicado en [38, 39].

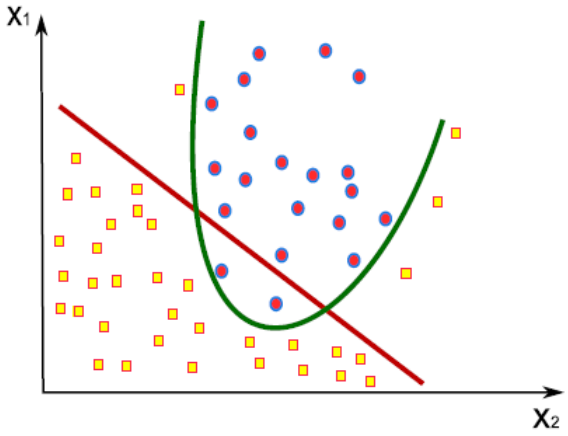

Figura 4.10: Clasificador No Lineal.

4.4.3. SVM con LBP

Este clasificador se utilizó de manera similar al punto anterior, con la diferencia que el vector de características utilizado como entrada al clasificador, se determinó con la aplicación y concatenación del algoritmo LBP [40].

4.4.3.1. LBP (*Local Binary Pattern*)

Las características LBP, son calculadas de acuerdo a la intensidad del pixel central $I(xc, yc)$ respecto de sus vecinos $I(x, y)$, tal que $h(I(xc, yc), I(x, y))$. Si $I(xc, yc) \leq I(x, y)$implica que $h = 1$, en caso contrario $h = 0$, Figura (4.11) luego:

$$LBP(x,y) = \cup_{(x',y')\epsilon N(x,y)} h(I(x,y),\ I(x',y')) \tag{4.9}$$

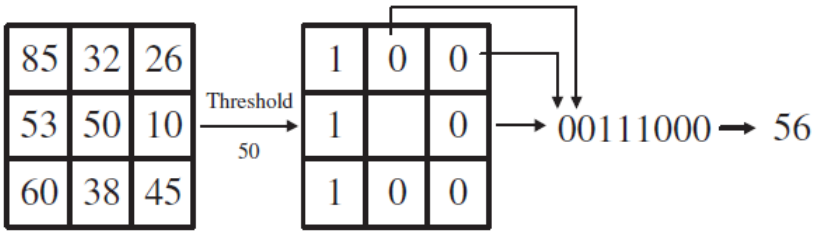

Figura 4.11: Operador Básico LBP, Fig. extraída de [3].

donde $N(x,y)$ es la vecindad alrededor de (x,y) y $\cup$ es el operador de concatenación.

4.4.4. ADABOOST

El algoritmo de Boosting Adaptivo es basado en un conjunto de clasificadores débiles en cascada, se utilizó de 2 maneras diferentes, primero como detector para encontrar el rostros presente en las imágenes, utilizando como clasificador débil las características Haar.

En una segunda etapa se utilizó como clasificador binario, para género mediante árboles de decisión (*decisión stumps*, en inglés) para lo cual se utilizó como vector de entrada, un vector de características extraídas en forma incremental desde las imágenes de rostros usando mRMR, NMIFS y CMIFS.

El número de características seleccionadas están en el rango de 50 a 500 para imágenes de 24x24, 50 a 1000 para imágenes de 36x36 y 50 a 2000 para imágenes de 48x48, el tamaño del vector se escogió empíricamente mediante pruebas. Diferentes implementaciones fueron usadas: Threshold, Real, Gentle [36], LUT [13] and Modest ADABOOST [41].

En el algoritmo original [36], los clasificadores débiles son construidos al comparar el valor de sus características con un umbral y de esa forma obtener una salida binaria. La característica y su salida son seleccionadas de acuerdo a la evaluación del error mínimo en cada etapa.

Algoritmo 4.1 Algoritmo ADABOOST tradicional.

- Dado un conjunto de imágenes (x_1, y_1),...,(x_n, y_n), donde y_i es 1 para los ejemplos positivos y 0 para los negativos .
- Inicialice los pesos $w_{1,i} = \frac{1}{2m}$, $\frac{1}{2l}$ para $y_i = 0, 1$ respectivamente, donde m y l son los ejemplos positivos y negativos.
- Para $t = 1, ...T$:
  1. Normalice los Pesos: $w_{t,i} \leftarrow \frac{w_{t,i}}{\sum_{j=1}^{n} w_{t,j}}$, tal que w_t es la probabilidad de distribución.
  2. Para cada característica, j se entrena un clasificador h_j, la cual es restringida a usar solo una característica.
     El error es evaluado con respecto a w_t, $\in_j = \sum_i w_i \mid h_j(x_i) - y_i \mid$.
  3. Escoja el clasificador h_t , con el menor error $\in_t$
  4. Actualice los pesos: $w_{t+1,\,i} = w_{t,i}\, \beta_t^{1-e_i}$, donde :
     a) $e_i = 0$, si el ejemplo x_i es clasificado correctamente.
     b) $e_i = 1$, de otra manera, y $\beta_t = \frac{\in_t}{1-\in_t}$.
  5. El clasificador fuerte, creado finalmente es:
     a) $h(x) = 1$, $\sum_{t=1}^{T} \alpha_t \cdot h_t(x) \geq \frac{1}{2} \sum_{t=1}^{T} \alpha_t$.
     b) $h(x) = 0$, de otra manera, donde $\alpha_t = log\frac{1}{\beta_t}$
- Fin

En ADABOOST del tipo umbral (*Threshold*), cada clasificador débil es seleccionado de acuerdo a un valor de umbral. Cuando una imagen es clasificada con el clasificador débil correspondiente, el valor calculado se compara con el umbral seleccionado, de acuerdo a esto se decide si es hombre o mujer.

El valor óptimo del umbral, es determinado durante el proceso de entrenamiento del algoritmo, de manera que el error sea el menor posible, para ese conjunto de datos.

En contraste Gentle ADABOOST (GAB), el cual basa el mecanismo de selección de clasificadores débiles, no forzando a obtener valores binarios, en vez de eso, obtienen un valor que se actualiza la función $f_m(x)$ a ejemplos de números finitos.

Algoritmo 4.2 Algoritmo Gentle ADABOOST.

1. Comience con pesos $w_i = 1/N, i = 1, ...N, F(x) = 0$.
2. Repita para $m = 1, 2, ...M$.
   a) Estimar $f_m(x)$ por el ajuste de pesos mediante mínimos cuadrados de y a x.
   b) Actualizar $F(x) \leftarrow F(x) + f_m(x)$.
   c) Colocar $w_i \leftarrow w_i exp\left[-y_i * f_m(x_i)\right], i = 1, 2, ...N$ y renormalizar los pesos $\sum_i w_i = 1$.
3. La salida del clasificador se obtiene de $sign\left[F(x)\right] = sign\left[\sum_{m=1}^{M} f_m(x)\right]$.

La principal diferencia entre Real ADABOOST (RAB) y Gentle ADABOOST (GAB) es la manera de estimar los pesos de las probabilidades de las clases y la actualización de los clasificadores débiles, $f_m(x)$. En GAB la actualización es dada por:

$$f_m(x) = P_w(y = 1 \mid x) \tag{4.10}$$

mientras que en RAB es dado por:

$$f_m(x) = \frac{1}{2} log \frac{P_w(y = 1 \mid x)}{P_w(y = -1 \mid x)} \tag{4.11}$$

En el caso de Modest ADABOOST [41], se indica que esta implementación tiene la capacidad de generalizar mejor y disminuir el error, dado por el siguiente pseudocódigo:

Algoritmo 4.3 Algoritmo Modest ADABOOST.

- Dado un conjunto de datos de entrenamiento $(x_1, y_1), ..., (x_n, y_n)$,inicialice los pesos de los datos $D_0(i) = 1/N$
- Para $m = 1, ..., M$ y mientras $f_m \neq 0$
- Entrene el clasificador débil $h_m(x)$ usando la distribución $D_m(i)$ por mínimos cuadrados.
- Calcule la distribución inversa $\overline{D_m}(i) = (1 - D_m(i))\overline{Z_m}$
- Calcule:
  - $P_m^{+1}(x) = P_{D_m}(y = +1 \cap h_m(x))$
  - $\overline{P}_m^{+1}(x) = P_{\overline{D}_m}(y = +1 \cap h_m(x))$
  - $P_m^{-1}(x) = P_{D_m}(y = -1 \cap h_m(x))$
  - $\overline{P}_m^{-1}(x) = P_{\overline{D}_m}(y = -1 \cap h_m(x))$
- Luego
  - $f_m(x) = (P_w^{+1}(1 - P_m^{+1}) - P_w^{-1}(1 - P_w^{-1}))(x)$
- Se actualiza la distribución:
  - $D_{m+1}(i) = \frac{D_m(i)\, exp\,(-y_i f_m(x_i))}{Z_m}$
- Construcción final del clasificador
  - $\sum_{i=1}^{i=M} f_{m(x)}$
- Fin

En el caso de LUT ADABOOST, es diferente, ya que en vez de calcular y comparar el valor con un umbral, se ubican dentro de un vector, el cual representa una cantidad de bines, definidos previamente en la etapa de entrenamiento. El rango de los valores de las características es dividido igualmente entre el número de bines seleccionados. Cuando la imagen de un rostro es clasificada, el valor de sus características es comparada con la almacenada en cada bin. Si el bin tiene más características femeninas, calculadas durante el entrenamiento que características masculinas, entonces es considerada femenina y viceversa.

En la práctica, el clasificador débil, es usado para construir la tabla LUT con la información de dichos clasificadores.

Supongamos una característica Haar $f_{haar}(x)$es normalizado a (0,1) y el rango es dividido en n sub-regiones, $bin = [(j-1)/n, j/n)]$ $j = 1, ..., n$.

Sea un ejemplo x , si $f_{Haar}(x) \in bin_j$, luego la predicción del clasificador débil, calculada

para está característica es:

$$h(x) = sign(P_1^{(j)} - P_2^{(j)}) \tag{4.12}$$

donde $P_i^{(j)} = P(x \in w_i \mid f_{Haar}(x) \in bin_j), i = 1, 2,\ j = 1, ..., n$ y $w_{1,}w_2$ representan las clases positivas y negativas.

$$h_{Lut}(x) = \sum_{j=1}^{n} sign(P_1^{(j)} - P_2^{(j)}) B_n^j(f_{Lut}(x)), \tag{4.13}$$

Donde,

$$B_n^j(u) = \{1 \in [j - 1/n, j/n], 0 \notin\},\ j = 1, .., n. \tag{4.14}$$

4.4.4.1. Clasificadores débiles

En este trabajo de investigación, se utilizaron árboles de decisión como clasificadores débiles para la implementación de ADABOOST. Un árbol de clasificación es un árbol gráfico, con hojas que representan los resultados de clasificación y los nodos que representan la predicción. Las hojas de los árboles están marcadas como verdaderos o falsos, como se aprecia en la Figura 4.12. En el proceso de clasificación se recorre el árbol de acuerdo a la profundidad definida en la configuración inicial, se comienza en la raíz y se sube hasta la última hoja, el valor asociado a esta última hoja representa la clase a predecir.

Este algoritmo fue construido de acuerdo al siguiente pseudocódigo:

Sea $S = \{(x^1, y^1), ..., (x^m, y^m)\}$ una secuencia de ejemplos de entrenamiento, donde x^jpertenece al dominio o espacio $X \in R^n$ (vector de valores reales con dimensionalidad n, $(x^j = (x_1^j, ..., x_n^j))$y cada etiqueta y^j pertenece al espacio finito de etiquetas Y, fue considerada clasificación binaria, en donde $Y = \{-1, 1\}$:

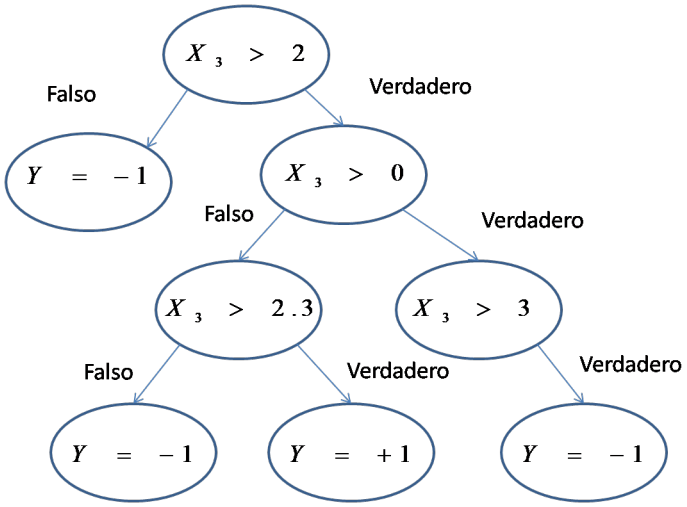

Figura 4.12: Ejemplo Gráfico de un árbol de decisión.

Algoritmo 4.4 Algoritmo básico de un árbol de clasificación.

1. Para cada una de las n dimensiones encuentre el umbral, que mejor separa S con el mínimo error.
2. Escoja la dimensión i con el menor error y construya el nodo:
   a) La predicción $x_i \geq umbral$
   b) La ramas verdaderas/falsas, que están conectadas con las hojas indicarán la clasificación.
3. Para hacer que el árbol aprenda los pesos de los datos de entrenamiento, solo debemos evaluar todos los errores de acuerdo a sus pesos.
4. Los datos de clasificación usan uniones de clasificadores débiles con respecto a su peso.
5. El resultado contiene números reales, en donde el signo representa la clase y la magnitud el valor de certeza de la decisión.

4.5. Datos

Cualquier método de clasificación usa un conjunto de patrones para caracterizar cada uno de los objetos a estudiar. Estos patrones deben ser relevantes para el objetivo de la clasificación. Se utilizan métodos llamados de clasificación supervisada donde un experto humano determina en que clases un objeto puede ser caracterizado. A-priori se necesita cierto conocimiento del objeto a clasificar, sin embargo, no necesita de conocimiento exacto sobre las diferencias que el clasificador ocupa para hacer las detecciones, es decir, se usa un aprendizaje estadístico.

Dentro de los métodos basados en las imágenes que ocupan aprendizajes estadísticos para construir sus patrones, existen distintas alternativas. Dentro de los más comúnmente usados están PCA (Análisis de Componentes Principales), SVM (Máquina de Soporte Vectorial), árboles de decisión, redes neuronales y ADABOOST. La mayor dificultad en usar estos métodos es la alta dimensionalidad del objeto a clasificar, por ejemplo: género en imágenes, donde cada imagen se representa como un vector debido a esto, un paso sumamente importante debe ser el reducir la dimensionalidad del problema y detectar aquellas características que mejor permiten diferenciar los rostros de hombre con el de una mujer.

Dos bases de datos internacionales fueron usadas para el entrenamiento y prueba de los métodos de selección de características, para poder comparar los resultados con las publicaciones previas. La base de datos FERET [14] la cual contiene imágenes frontales en escala de grises de 1199 individuos con diferentes poses e iluminación uniforme. Como en Makinen y Raisano (2008b), se utilizó solo 1 rostro por persona desde los subconjuntos Fa y Fb en donde las imágenes duplicadas fueron eliminadas. De esta manera se seleccionaron 450 mujeres y 450 hombres.

La segunda Base de datos Utilizada fue la WEB [10]. Para poder comparar nuestros resultados con [3], 2 tamaños de imágenes fueron utilizados: 24x24 y 32x40, un ejemplo de imágenes de cada base de datos puede ser apreciado en la Figura 4.13. Se compararon además los resultados con [3] en donde se utilizaron 304 imágenes y 3 tamaños de imágenes (24x24, 36x36 y 48x48) desde el subconjunto Fa de FERET. Las imágenes se encuentran disponibles en el link del autor [3].

Para los cuatro modelos, el conjunto de entrenamiento fue utilizado para determinar el mejor número de características seleccionadas, para luego ser evaluadas con el conjunto de prueba. La división de ambos conjuntos se realizó por única vez, de forma aleatoria determinando un 80 % para entrenamiento y 20 % para prueba. Todos los resultados fueron obtenidos con 5 simulaciones, utilizando validación cruzada de 5 grupos.

Figura 4.13: Ejemplos de imágenes utilizadas, izquierda Base de datos FERET, derecha bases de datos Web.

Capítulo 5

Resultados

La tabla 5.1 compara los resultados previamente publicados para los clasificadores SVM, NN y ADABOOST para diferentes tamaños de imágenes, 24x24, 36x36 y 48x48, sobre la base de datos FERET con nuestros resultados utilizando Información Mutua y cuatro clasificadores diferentes. Los resultados representan el promedio de 5 simulaciones en validación cruzada de 5 grupos diferentes con una partición aleatoria de imágenes de las bases de datos, realizada por única vez, con la finalidad de poder comparar los resultados de los diferentes clasificadores.

Las primeras cuatro filas de la tabla 5.1 muestran el resultado de la mejor tasa de clasificación publicada en Makinen y Raisano (2008a) para los clasificadores SVM, SVM_LBP, NN y ADABOOST Threshold, para 3 tamaños de imágenes diferentes: 24x24, 36x36, 48x48 pixeles. Cada columna muestra entre paréntesis el número de características seleccionadas con la que se obtuvo el máximo de clasificación en el conjunto de prueba y la desviación estándar para cada modelo. Las filas desde la 5 a la 32 muestran los resultados de los mismos métodos de clasificación pero utilizando la selección de características, mRMR (MID y MIQ), NMIFS, CMIFS. El mejor resultado de reconocimiento correcto de rostros fue de 94.3 % en la base de datos FERET para tamaño de rostro 24x24, para el clasificador Modest ADABOOST y método de selección CMIFS para 400 características. Este resultado es 9.6 % más alto que el mejor resultado previamente publicado con 576 características. Para las imágenes de tamaño 36x36 de la base de datos FERET, el mejor resultado fue de 95.6 % de clasificación correcta de rostros utilizando el algoritmo Real ADABOOST y el método de selección NMIFS para 850 características. Este resultado es 8.6 % más alto que los resultados previamente publicados con 1296 características utilizando SVM. Para el caso de las imágenes de 48x48 de la base de datos FERET el mejor resultado fue de 96.78 % alcanzado con SVM_LBP utilizando 350 característica con NMIFS, lo cual es mucho mejor a lo previamente publicado y con un número significativamente menor de características.

Tabla 5.1: Resultados de la aplicación de algoritmos de selección de características a imágenes FERET, se indica su desviación estándar y N° de características seleccionadas en paréntesis.

Método	FERET 24x24 (%)	FERET 36x36 (%)	FERET 48x48(%)
SVM [3]	82.64± 0.593(576)	86.32± 4.257(1296)	84.12± 4.939(2304)
SVM_LBP [3]	76.90± 0.806(576)	80.00± 0.931(1296)	83.01± 0.618(2304)
NN [3]	84.37± 0.505(576)	86.97± 1.999(1296)	81.96± 5.863(2304)
ADA_TH [3]	82.35± 1.107(576)	84.25± 0.672(1296)	86.90± 1.189(2304)
SVM_MID	92.54± 2.196(350)	93.17± 1.868(300)	90.69± 1.285(850)
SVM_MIQ	91.72± 0.831(300)	91.26± 2.066(350)	89.19± 2.369(800)
SVM_NMIFS	92.53± 1.408(250)	91.89± 1.203(900)	89.83± 0.789(1000)
SVM_CMIFS	91.19± 2.982(450)	92.48± 1.877(600)	95.51± 1.148(750)
SVM_LBP_MID	90.15± 1.234(150)	92.07± 0.753(300)	92.10± 0.763(350)
SVM_LBP_MIQ	90.20± 0.736(150)	91.63± 0.787(300)	91.69± 0.539(350)
SVM_LBP_NMIFS	91.45± 0.741(100)	90.68± 1.188(100)	**96.78± 1.504(350)**
SVM_LBP_CMIFS	89.84± 0.990(200)	90.11± 1.292(100)	94.08± 1.642(350)
NN_MID	86.85± 0.791(300)	90.02± 2.618(550)	87.32± 1.019(800)
NN_MIQ	89.83± 0.653(300)	90.02± 2.618(550)	89.49± 0.961(800)
NN_NMIFS	89.83± 0.635(300)	89.94± 0.743(100)	89.49± 0.961(800)
NN_CMIFS	90.15± 0.589(400)	90.42± 0.618(350)	90.04± 1.197(800)
ADA_TH_MID	86.21± 1.107(250)	89.33± 0.672(350)	87.55± 1.189(900)
ADA_TH_MIQ	82.28± 0.514(250)	87.35± 1.972(350)	87.37± 1.565(850)
ADA_TH_NMIFS	91.21± 0.644(400)	92.18± 0.984(800)	92.62± 1.177(600)
ADA_TH_CMIFS	90.61± 0.709(400)	91.46± 0.761(600)	91.41± 0.931(800)
ADA_REAL_MID	88.83± 1.346(200)	90.04± 1.558(250)	86.53± 1.349(850)
ADA_REAL_MIQ	87.14± 0.505(250)	89.33± 1.999(300)	85.74± 5.863(900)
ADA_REAL_NMIFS	93.52± 0.750(500)	**95.58± 1.934(850)**	84.24± 4.676(500)
ADA_REAL_CMIFS	90.19± 1.254(550)	94.20± 2.273(900)	84.92± 5.893(200)
ADA_GENTLE_MID	89.85± 0.806(200)	89.54± 3.022(250)	87.97± 4.576(900)
ADA_GENTLE_MIQ	88.56± 0.662(300)	88.47± 3.850(400)	85.76± 3.651(850)
ADA_GENTLE_NMIFS	93.30± 0.718(400)	93.28± 2.046(400)	93.57± 1.658(200)
ADA_GENTLE_CMIFS	93.28± 0.671(150)	93.24± 0.728(950)	94.34± 1.583(200)
ADA_MOD_MID	91.55± 1.367(200)	89.53± 2.211(250)	86.78± 5.016(900)
ADA_MOD_MIQ	90.03± 0.593(250)	89.72± 4.257(300)	86.27± 4.939(1000)
ADA_MOD_NMIFS	93.86± 0.794(150)	94.83± 1.872(750)	91.70± 1.324(350)
ADA_MOD_CMIFS	**94.30± 0.918(400)**	93.13± 1.493(900)	94.47± 2.567(500)

La tabla 5.2 muestra en las 4 primeras filas, los mejores resultados publicados para la clasificación de género en la base de datos FERET y WEB para 2 tamaños diferentes de rostros, 24x24 y 32x40 pixeles para los clasificadores SVM, SVM_LBP, NN y LUT.

Los resultados son el promedio de 5 simulaciones con partición aleatoria de datos en

validación cruzada en 5 grupos, considerando 80 % para entrenamiento y 20 % para prueba. Las primeras cuatro columnas de la tabla 5.2 muestran los resultados de lo publicado en Makinen y Raisano (2008b). Las filas de la 5 a la 20 muestran los mismos clasificadores pero usando los métodos de selección de características propuestos en este trabajo, mRMR (MID y MIQ), NMIFS, y CMIFS. Cada columna muestra el promedio de 5 simulaciones, la desviación estándar y en paréntesis el número de las características seleccionadas en cada modelo. El mejor resultado fue de 94.08 % de clasificaciones correctas de rostros en la base de datos FERET para tamaño 24x24 pixeles, el mejor clasificador resulto ser SVM y el método de selección NMIFS con 400 características. Este resultado fue 13.12 % más alto que el previamente publicado con 576 características (Primera fila de la tabla 5.2).

El mejor resultado 94.41 % para las imágenes de rostros de 32x40 pixeles fue obtenida con un clasificador SVM y método de selección de características NMIFS con un total de 950 características. Este resultado fue 1.2 % mayor que los resultados previamente publicados con 1280 (32x40) características usando LUT ADABOOST.

El mejor resultado para la base de datos WEB, fue de 83.86 % de clasificaciones correctas para imágenes de rostros de tamaño 24x24 pixeles. Este resultado fue obtenido por un clasificador SVM_LBP con método de selección MID con un total de 150 características. Este resultado fue 1.1 % más alto que el mejor resultado publicado con 576 características (primera fila de la tabla 5.2).

El mejor resultado para la base de datos WEB con imágenes de rostros de tamaño 32x40 fue de 86 % de clasificaciones correctas obtenido con SVM_LBP, método de selección NMIFS con 150 características. Este resultado fue 8.9 % más alto que lo previamente publicado con 1280 (32x40) características usando SVM_LBP. En resumen, los métodos propuestos de selección de características reducen y mejoran significativamente la clasificación en 8.9 % sobre las comparaciones realizadas en la base WEB, con lo previamente publicado.

Además de mejorar las resultados, los métodos de selección de características reducen el número de características desde 576 (24x24) y 1280 (32x40) hasta 400 características en la base de datos FERET y 150 características en la base de datos WEB. Por lo tanto, el tiempo computacional es reducido de manera importante para sus implementaciones en tiempo real en 69.4 % para la base de datos FERET en imágenes de tamaño 24x24 pixeles y en 74.2 % con imágenes 32x40 pixeles.

En el caso de la base de datos WEB las características pueden ser reducidas a 26 % para imágenes 24x24 y en 11.7 % para imágenes de 32x40. El tiempo computacional puede ser muy importante comercialmente, si el método de clasificación de género es utilizado en tiempo real, como por ejemplo publicidad electrónica en tiendas de retail o marketing selectivo. Por lo tanto la selección de característica es altamente deseable.

Tabla 5.2: Resultados de la aplicación de algoritmos de selección de características a imágenes FERET y Web, se indica su desviación estándar y N° de características en paréntesis.

Método	FERET 24x24 (%)	FERET 32x40 (%)	WEB 24x24 (%)	WEB 32x40 (%)
SVM [10]	87.15±0.102 (576)	81.29±0.111 (1280)	79.74±0.077 (576)	75.77±0.077 (1280)
SVM_LBP [10]	81.12±0.155 (576)	91.80±0.107 (1280)	73.84±0.095 (576)	77.07±0.086 (1280)
NN [10]	91.79±0.107 (576)	90.16±0.071 (1280)	72.61±0.066 (576)	61.94±0.070 (1280)
LUT_Ada [10]	89.89±0.117 (576)	93.24±0.050 (1280)	74.47±0.096 (576)	76.63±0.084 (1280)
SVM_MID	94.00±0.006 (200)	94.26±0.020 (800)	81.09±0.036 (400)	79.89±0.017 (500)
SVM_MIQ	93.30±0.018 (250)	93.86±0.011 (900)	79.74±0.026 (400)	79.22±0.021 (500)
SVM_NMIFS	**94.08±0.022 (400)**	**94.41±0.015 (950)**	79.95±0.014 (400)	80.33±0.045 (900)
SVM_CMIFS	93.26±0.023 (350)	93.22±0.011 (850)	81.83±0.036 (550)	76.67±0.046 (650)
SVM_LBP_MID	89.69±0.012 (150)	92.68±0.012 (300)	**83.86±0.027 (150)**	83.09±0.019 (300)
SVM_LBP_MIQ	90.27±0.011 (150)	92.36±0.011 (300)	82.40±0.026 (200)	81.15±0.018 (300)
SVM_LBP_NMIFS	86.28±0.007 (150)	90.59±0.011 (300)	78.71±0.019 (200)	**86.00±0.017 (150)**
SVM_LBP_CMIFS	91.00±0.015 (150)	92.28±0.010 (300)	79.39±0.025 (200)	80.42±0.013 (300)
NN_MID	91.57±0.551 (200)	91.22±0.101 (250)	79.11±0.080 (300)	70.43±0.045 (500)
NN_MIQ	89.52±0.627 (250)	90.29±0.099 (300)	78.37±0.081 (350)	70.71±0.102 (500)
NN_NMIFS	89.39±0.542 (450)	90.47±0.121 (450)	79.83±0.091 (450)	80.26±0.051 (450)
NN_CMIFS	89.52±0.839 (400)	91.17±0.105 (400)	75.12±0.106 (450)	77.05±0.050 (500)

Tabla 5.2: Continuación

Método	FERET 24x24 (%)	FERET 32x40 (%)	WEB 24x24 (%)	WEB 32x40 (%)
LUT_Ada_MID	89.51±0.802 (400)	92.87±0.143 (350)	78.17±0.053 (400)	79.40±0.067 (350)
LUT_Ada_MIQ	89.40±1.116 (450)	90.67±0.103 (450)	77.10±0.097 (400)	78.36±0.070 (500)
LUT_Ada_NMIFS	90.41±0.797 (350)	91.43±0.143 (500)	77.03±0.090 (350)	77.11±0.057 (500)
LUT_Ada_CMIFS	90.05±1.069 (350)	91.26±0.087 (500)	77.21±0.066 (350)	76.49±0.054 (500)

Las figuras 5.1 a la 5.12 muestra ejemplos de los métodos de selección de características para los métodos MID, NMIFS y CMIFS. Estas figuras muestran resultados y distribución de características para imágenes de hombre y mujer, desde la base datos FERET y WEB.

Las figuras 5.1 a la 5.12 además muestran las características seleccionadas desde 100 a 1000 con los diferentes métodos para la base de datos FERET y WEB. En ellas se puede apreciar como se distribuyen sobre el rostro y sus alrededores, lo importante es poder escoger el método que mejor seleccione la mayor cantidad de pixeles dentro del rostro y sus bordes. En el caso del método CMIFS la selección se lleva a cabo desde los bordes hacia adentro, se aprecia que la zona de la frente del rostro no entrega mucha información, a diferencia de mRMR y NMIFS en donde la selección comienza desde el interior del rostro hacia afuera, se debe tener presente que cuando 2 pixeles son iguales su información mutua es cero y no se seleccionan.

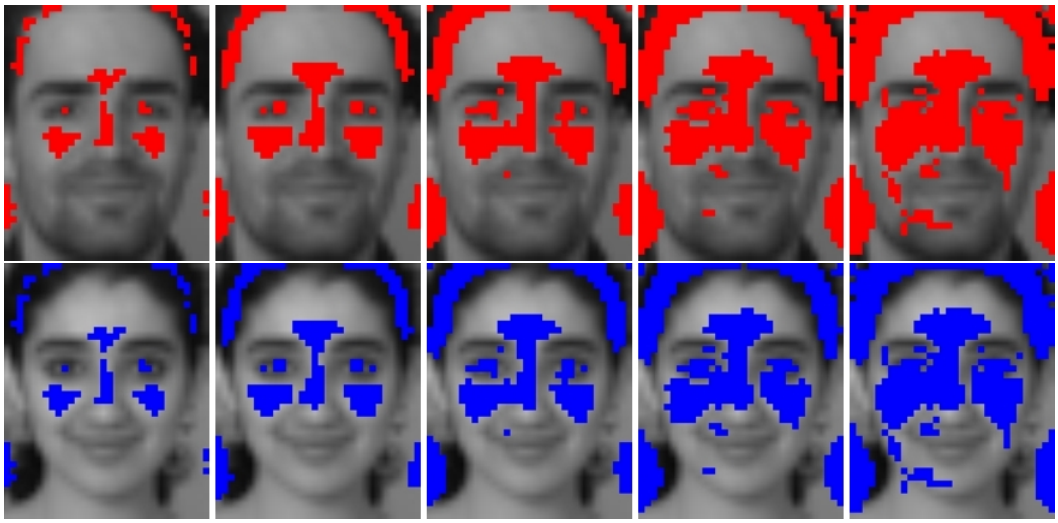

Figura 5.1: Imágenes originales de tamaño 32x40, un hombre y una mujer, para la base de datos FERET subconjunto Fa con 100 a 500 características seleccionadas para la clasificación de género, con el algoritmo mRMR.

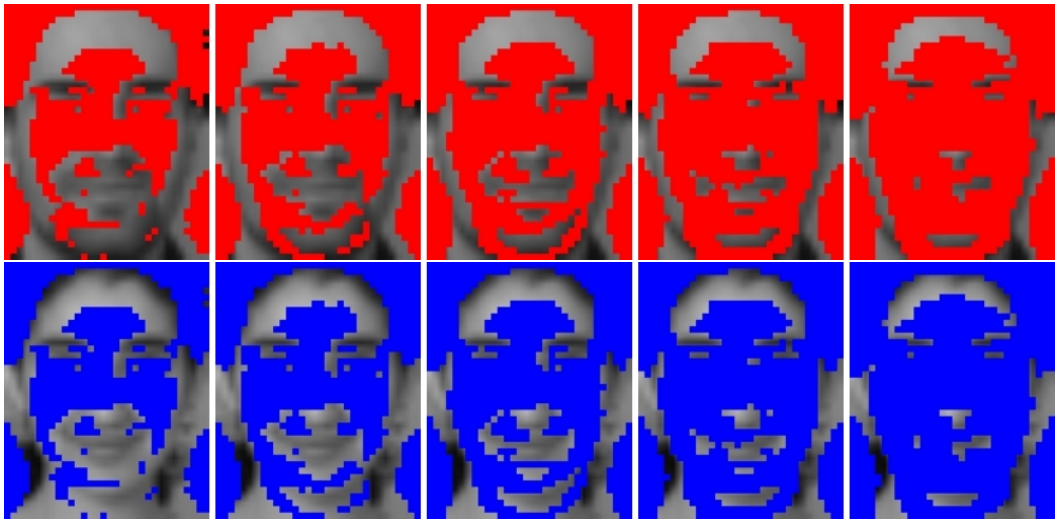

Figura 5.2: Imágenes originales de tamaño 32x40, un hombre y una mujer, para la base de datos FERET subconjunto Fa con 600 a 1000 características seleccionadas para la clasificación de género, con el algoritmo mRMR.

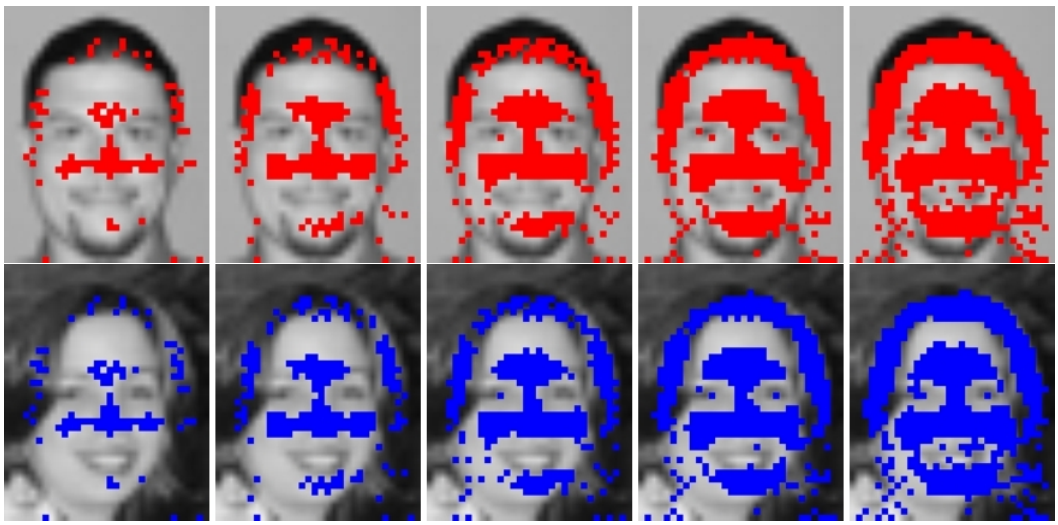

Figura 5.3: Imágenes originales de tamaño 32x40, un hombre y una mujer, para la base de datos WEB con 100 a 500 características seleccionadas para la clasificación de género, con el algoritmo mRMR.

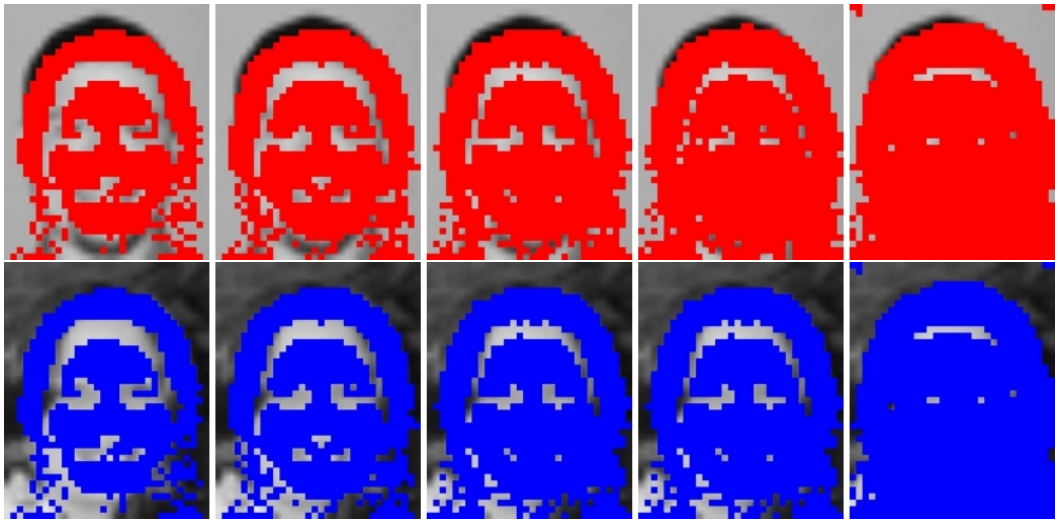

Figura 5.4: Imágenes originales de tamaño 32x40, un hombre y una mujer, para la base de datos WEB con 600 a 1000 características seleccionadas para la clasificación de género, con el algoritmo mRMR.

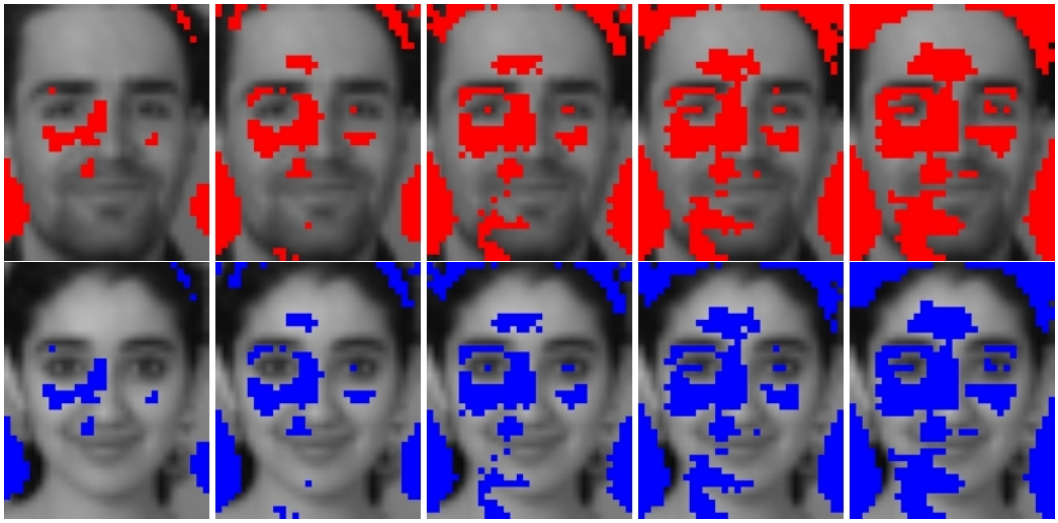

Figura 5.5: Imágenes originales de tamaño 32x40, un hombre y una mujer, para la base de datos FERET subconjunto Fa con 100 a 500 características seleccionadas para la clasificación de género, con el algoritmo NMIFS.

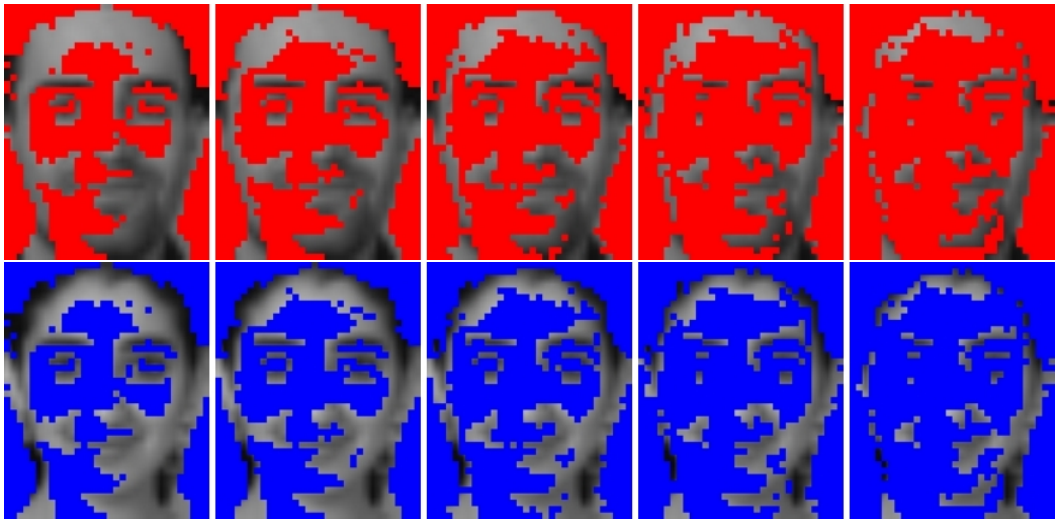

Figura 5.6: Imágenes originales de tamaño 32x40, un hombre y una mujer, para la base de datos FERET subconjunto Fa con 600 a 1000 características seleccionadas para la clasificación de género, con el algoritmo NMIFS.

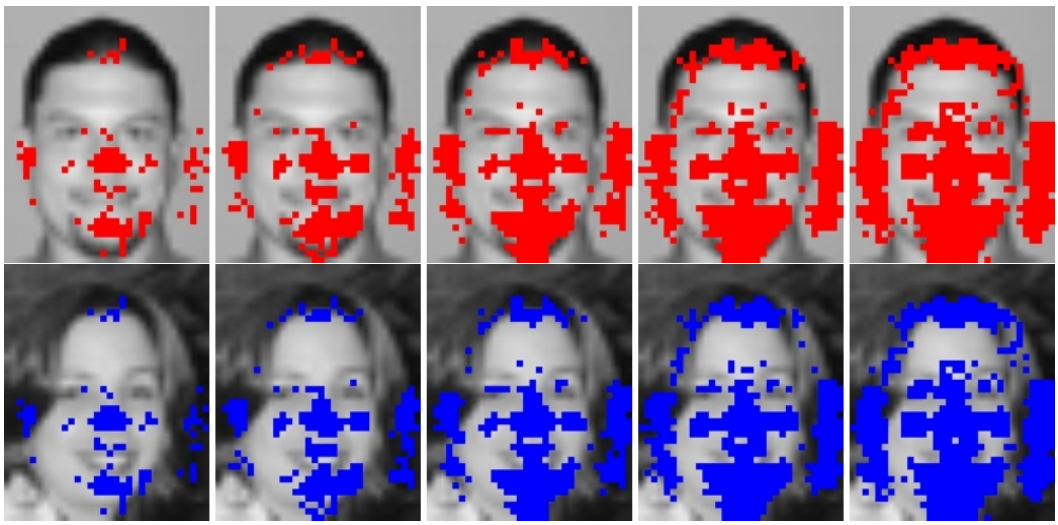

Figura 5.7: Imágenes originales de tamaño 32x40, un hombre y una mujer, para la base de datos Web con 100 a 500 características seleccionadas para la clasificación de género, con el algoritmo NMIFS.

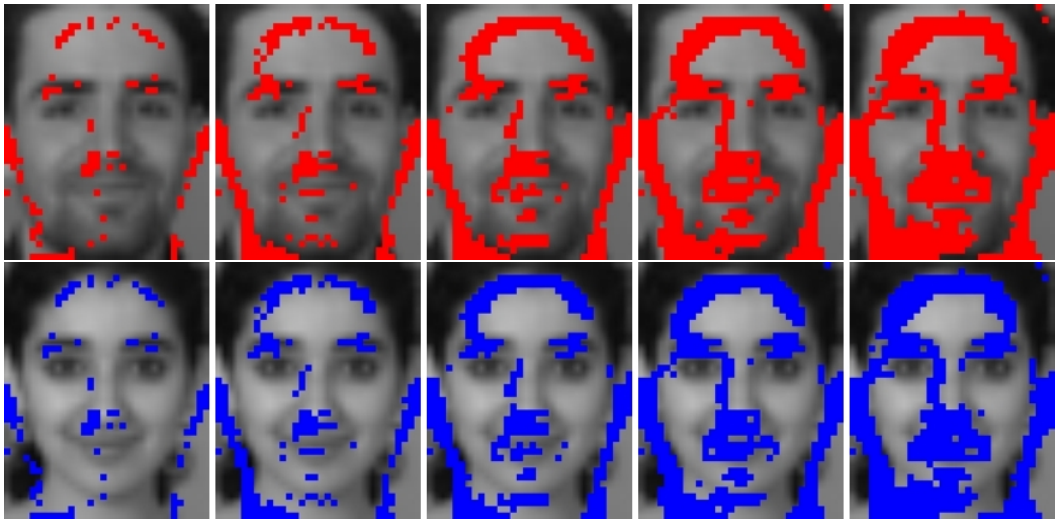

Figura 5.9: Imágenes originales de tamaño 32x40, un hombre y una mujer, para la base de datos FERET subconjunto Fa con 100 a 500 características seleccionadas para la clasificación de género, con el algoritmo CMIFS.

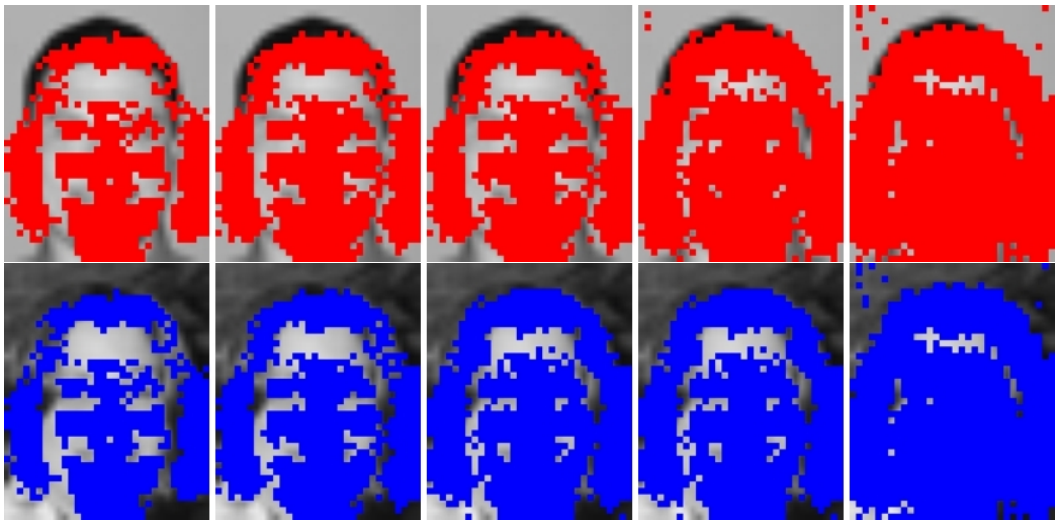

Figura 5.8: Imágenes originales de tamaño 32x40, un hombre y una mujer, para la base de datos Web con 600 a 1000 características seleccionadas para la clasificación de género, con el algoritmo NMIFS.

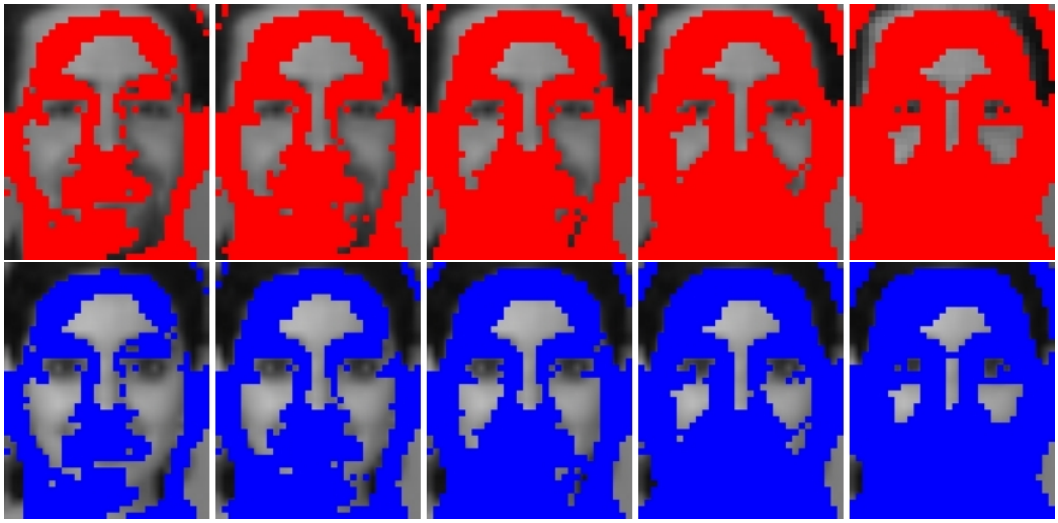

Figura 5.10: Imágenes originales de tamaño 32x40, un hombre y una mujer, para la base de datos FERET subconjunto Fa con 600 a 1000 características seleccionadas para la clasificación de género, con el algoritmo CMIFS.

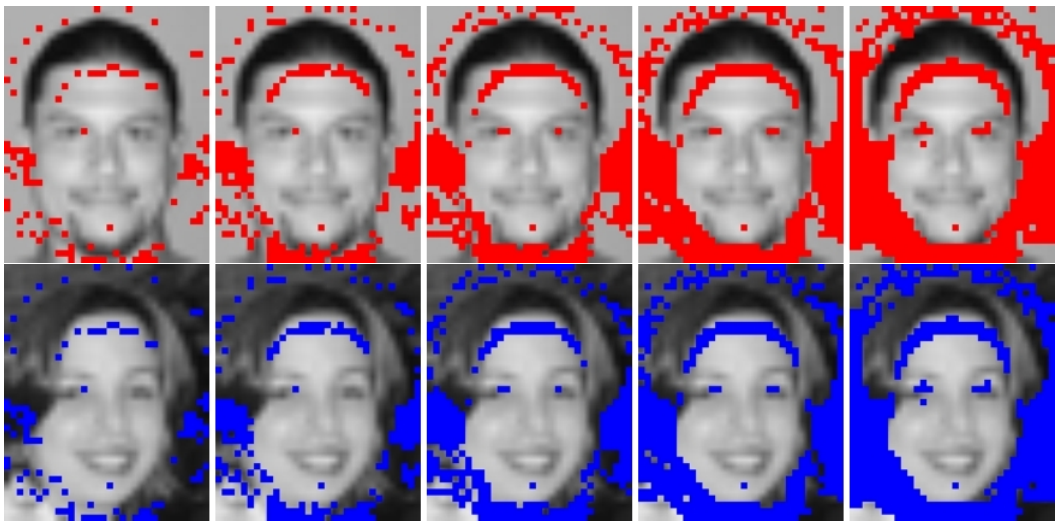

Figura 5.11: Imágenes originales de tamaño 32x40, un hombre y una mujer, para la base de datos Web con 100 a 500 características seleccionadas para la clasificación de género, con el algoritmo CMIFS.

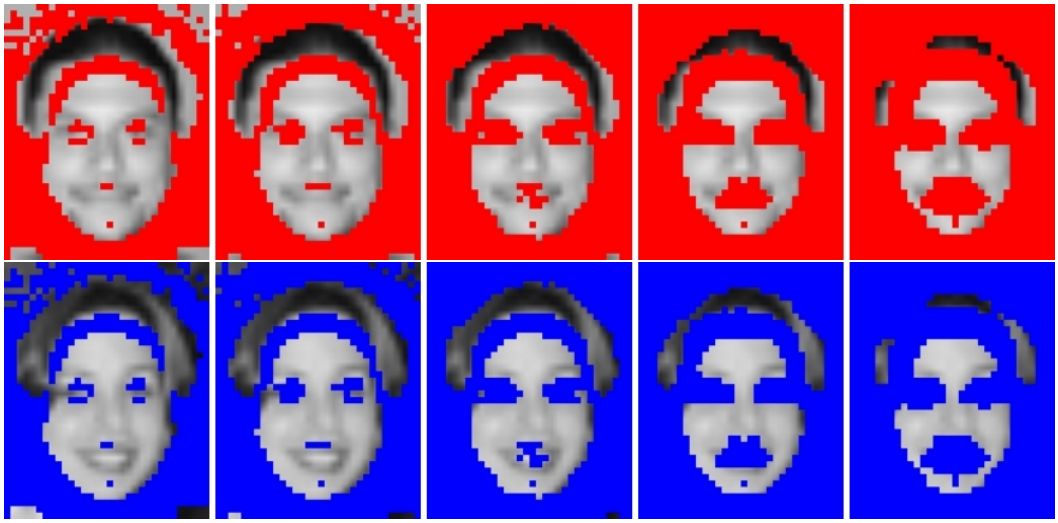

Figura 5.12: Imágenes originales de tamaño 32x40, un hombre y una mujer, para la base de datos Web con 500 a 1000 características seleccionadas para la clasificación de género, con el algoritmo CMIFS.

Las tablas 5.4 a 5.7, comparan el tiempo computacional para los clasificadores: SVM, SVM_LBP, NN y ADABOOST usando todas las características, comparado con los diferentes métodos de selección. Además de mejoras los resultados de clasificación de género, se aprecia que los métodos propuestos reducen el número de características desde 576 pixeles para imágenes (24x24) y 1280 para imágenes (32x40) a cerca de 400 en la base de datos FERET y cerca de 150 características en la base de datos WEB. De esta manera, el tiempo computacional para clasificar género en aplicaciones de tiempo real es reducido significativamente en 69.4 % en FERET para imágenes 24x24 y 74.2 % en FERET con imágenes de tamaño 32x40. En el caso de base de datos WEB esta razón puede ser reducida a un 26 % para imágenes de tamaño 24x24 y 11.7 % para imágenes de tamaño 32x40. El tiempo computacional es de importante interés comercial en la clasificación de género, ya que los métodos de selección de características pueden ser aplicados en tiempo real, por ejemplo en marketing electrónico para tiendas de retail.

Las tablas 5.4 y 5.5 muestran el tiempo computacional para los diferentes métodos de selección de características para la base de datos FERET, para imágenes alineadas y de iluminación controlada de tamaño 24x24 y 32x40. El tiempo de clasificación indica el tiempo consumido para clasificar la totalidad de las imágenes para cada clasificador, además de mostrar el tiempo que toma 1 imagen, luego de seleccionar sus características en ser clasificada. N/A, indica que no se realizó selección de características, utilizando todos las características para clasificar.

Todas las pruebas fueron realizadas en un PC Intel QuaD Core de 2.8 GHz de 4Gb de memoria Ram, utilizando el Software Matlab 7.8 (R2009a).

Tabla 5.4: Tiempo computacional consumido por los mejores métodos de selección de características, para la base de datos FERET 24x24.

FERET 24x24						
Clasificador	Resultado (%)	Método	Nº Imágenes	Nº Características	Tiempo Clasificación (seg)	Tiempo por Imagen (seg)
SVM	94.08	NMIFS	180	400	0.2401	0.00060
SVM-LBP	91.00	CMIFS	180	150	0.2166	0.00144
NN	91.57	CMIFS	180	200	0.1945	0.00097
ADA-LUT	90.41	NMIFS	180	350	0.2245	0.00064
SVM	87.15	N/A	180	576	0.4975	0.00276

Tabla 5.5: Tiempo computacional consumido por los mejores métodos de selección de características, para la base de datos FERET con imágenes de tamaño 32x40.

FERET 32x40						
Clasificador	Resultado (%)	Método	Nº Imágenes	Nº Características	Tiempo Clasificación (seg)	Tiempo por Imagen (seg)
SVM	94.41	NMIFS	152	950	0.61	0.00401
SVM-LBP	92.68	MID	152	300	0.50	0.00328
NN	93.22	MID	152	250	0.67	0.00440
ADA-LUT	92.87	MID	152	350	0.59	0.00388
SVM	81.29	N/A	152	1240	1.19	0.00782

Las tablas 5.6 y 5.7 muestran el tiempo computacional para los diferentes métodos de selección de características para la base de datos WEB, para imágenes no alineadas en diferentes ambientes y condición de iluminación no controlada de tamaño 24x24 y 32x40. El tiempo de clasificación indica el tiempo consumido para clasificar la totalidad de las imágenes para cada clasificador, además de mostrar el tiempo que toma 1 imagen, luego de seleccionar sus características en ser clasificada. N/A, indica que no se realizó selección de características, utilizando todos las características.

Tabla 5.6: Tiempo computacional consumido por los mejores métodos de selección de características, para la base de datos WEB con imágenes de tamaño 24x24.

WEB 24x24						
Clasificador	Resultado	Método	Nº	Nº	Tiempo	Tiempo
	(%)		Imágenes	Características	Clasificación	por Imagen
					(seg)	(seg)
SVM	78.71	CMIFS	944	550	0.310	0.00032
SVM-LBP	83.86	MID	944	150	0.183	0.00019
NN	79.83	NMIFS	944	450	0.294	0.00031
ADA-LUT	82.00	CMIFS	944	500	0.340	0.00036
SVM	79.74	S/S	944	576	0.760	0.00081

Tabla 5.7: Tiempo computacional consumido por los mejores métodos de selección de características, para la base de datos WEB con imágenes de tamaño 32x40.

WEB 32x40						
Clasificador	Resultado (%)	Método	Nº Imágenes	Nº Características	Tiempo Clasificación (seg)	Tiempo por Imagen (seg)
SVM	80.33	NMIFS	762	900	1.58	0.00207
SVM-LBP	83.09	MID	762	300	1.31	0.00172
NN	80.26	NMIFS	762	450	1.35	0.00177
ADA-LUT	79.40	MID	762	350	0.98	0.00128
SVM	75.77	N/A	762	1280	2.41	0.00316

5.1. Análisis estadístico

Se utilizó el test estadístico de ANOVA *(Analysis of Variance*, en inglés) [42] para determinar si existen diferencias significativas entre varias poblaciones o grupos. Del mismo modo que la t de Student, la prueba ANOVA es una prueba paramétrica y como tal requiere una serie de supuestos para poder ser aplicada correctamente. En realidad nos va a servir no solo para estudiar las dispersiones o varianzas de los grupos, sino para estudiar sus medias y la posibilidad de crear subconjuntos de grupos con medias iguales. Se puede decir que la prueba ANOVA es la generalización de la t de Student, ya que si realizamos una prueba ANOVA en la comparación de solo dos grupos, obtenemos los mismos resultados.

Al igual que la t de Student [42], se requiere que cada uno de los grupos a comparar tenga distribuciones normales, o lo que es más exacto, que lo sean sus residuales. Los residuales son las diferencias entre cada valor y la media de su grupo. Además debemos estudiar la dispersión o varianzas de los grupos, es decir estudiar su homogeneidad. Cuando mayor sean los tamaños de los grupos, menos importante es asegurar estos dos supuestos, ya que el ANOVA

suele ser una técnica bastante "robusta" comportándose bien respecto a transgresiones de la normalidad. El test estadístico del ANOVA es la razón entre dos medidas de variación de los datos muéstrales. El valor p será la probabilidad de observar un test estadístico tan o más grande.

Vamos a llamar factor p a una variable cualitativa que usaremos para designar a los grupos o tratamientos a comparar. Si el valor $p < 0{,}05$ quiere decir que no hay diferencias estadísticamente importantes entre los promedios, por lo tanto el metodo puede ser considerado robusto. Cuando se asume como nivel de significación el valor de 0,05 quiere decir que está dispuesto a asumir un riesgo de equivocarse de hasta el 5 % de las veces y decir que los dos grupos son diferentes cuando en realidad son iguales.

Se compararon los resultados de clasificación sin selección de características versus los resultados de clasificación con los diferentes métodos de selección de características, para cada tabla de resultados. En la tabla 1, para la base de datos FERET con imágenes de tamaño 24x24, 36x36 y 48x48, el test de ANOVA índico que los clasificadores: SVM-NMIFS, SVM-CMIFS, SVM-LBP, ADA-GENTLE-NMIFS, ADA-GENTLE-CMIFS, ADA-MOD-NMIFS y ADA-MOD-CMIFS, tienen medias significativamente diferentes. En todos estos casos el factor p el cual debe ser ($p < 0{,}05$), fue 1.27e-5 lo cual es altamente significativo estadísticamente. El mejor resultado fue obtenido con ADA-MODEST-CMIFS con 94.30 +/- 0.918 y 400 características. Este es el mejor caso en comparación a los 4 métodos de clasificación sin selección de características (Tabla 1, líneas 1 a 4).

En la Tabla 2, para la base de datos FERET, también se compararon los resultados de todos los clasificadores sin selección versus aquellos con selección de características con información mutua. El test de multicomparación ANOVA obtuvo un factor p=0.047. El mejor resultado fue el método de selección de características NMIFS con el clasificador SVM con 400 características. Todas las variantes de SVM con métodos de selección MID, MIQ, NMIFS y CMIFS mostraron resultados significativamente diferentes a los obtenidos con los mismos clasificadores pero sin selección (Tabla 2, líneas 1 a 4).

En el Caso de la base de datos WWW, también se obtuvieron resultados significativamente diferentes estadísticamente, cuando fueron comparados los resultados de clasificación con selección (MID, MIQ, NMIFS Y CMIFS) y sin selección de características. El test ANOVA alcanzó un valor p=0.0158, en donde la red neuronal alcanzó la tasa de clasificación mas baja. El mejor resultado fue obtenido con SVM-LBP-MID alcanzando un 86.00 % +/- 0.017 y 150 características seleccionadas.

Luego de analizar los resultados, se puede concluir que la selección de características mejora significativamente el resultado de la clasificación de género. La calidad de las imágenes en la base de datos FERET son muy superior a las ofrecidas por la base de datos WWW, lo cual indudablemente influye en los resultados.

5.2. Conclusión

Un nuevo método para clasificación de género es propuesto usando información mutua para seleccionar características faciales. Dos formas de combinar relevancia y redundancia usando MID y MIQ fueron empleados (mRMR) así como NMIFS y CMIFS. Los resultados muestran que la clasificación de género puede ser mejorada significativamente , cerca de 12.7 % con los métodos de selección de características.

Es relevante destacar que la disminución lograda en el número de características necesarias para poder clasificar un rostro. Reducir el número de características tiene el beneficio adicional de reducir los requerimientos computacionales, haciendo posible la implementación del método en tiempo real.

El mejor rendimiento en resultado de clasificación fue obtenido con NMIFS y muestra mejoras de 12.7 % para la base de datos FERET y 11.7 % sobre la base de datos WEB, comparados con los resultados previamente publicados.

5.2.1. Trabajo futuro

Considerando los resultados obtenidos en nuestro trabajo se puede indicar que, es posible mejorar los resultados, para lo cual:

- Trabajar en la Fusión de características a diferentes escalas de resolución, considerando los buenos resultados obtenidos con LBP, agregando otras características como por ejemplo la forma.
- Profundizar, en las diferentes variantes de LBP, en conjunto con la selección de características.
- Trabajar en clasificaciones multi-vistas extendiendo el estudio de imágenes frontales
- Trabajar en la implementación de una interfaz gráfica que permita probar estos avances en secuencias de videos y utilizar estos tópicos en la clasificación en tiempo real como un paso más de la detección y reconocimiento.

Referencias

[1] K. Irick, M. DeBole, V. Narayanan, R. Sharma, H. Moon, and S. Mummareddy, "A unified streaming architecture for real time face detection and gender classification," pp. 267–272, 2007.

[2] L. Lu, Z. Xu, and P. Shi, "Gender classification of facial images based on multiple facial regions," vol. 6, pp. 48–52, 2009.

[3] E. Mäkinen and R. Raisamo, "Evaluation of gender classification methods with automatically detected and aligned faces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 541–547, 2008a.

[4] Y. Wang, H. Ai, B. Wu, and C. Huang, "Real time facial expression recognition with adaboost," vol. 3, pp. 926–929, 2004.

[5] J. Wu, W. A. Smith, and E. R. Hancock, "Facial gender classification using shape-from-shading," *Image and Vision Computing*, vol. 28, no. 6, pp. 1039 – 1048, 2010.

[6] L. A. Alexandre, "Gender recognition: A multiscale decision fusion approach," *Pattern Recognition Letters*, vol. 31, no. 11, pp. 1422 – 1427, 2010.

[7] C. Perez, V. Lazcano, P. Estevez, and C. Estevez, "Real-time iris detection on faces with coronal axis rotation," *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, vol. 7, pp. 6389 – 6394 vol.7, 2004.

[8] C. Perez, L. Castillo, L. Cament, P. Estevez, and C. Held, "Genetic optimisation of illumination compensation methods in cascade for face recognition," *Electronics Letters*, vol. 46, no. 7, pp. 498–500, 2010.

[9] C. A. Perez, C. M. Aravena, J. I. Vallejos, P. A. Estevez, and C. M. Held, "Face and iris localization using templates designed by particle swarm optimization," *Pattern Recognition Letters*, vol. 31, no. 9, pp. 857 – 868, 2010.

[10] Mäkinen and Raisano, "An experimental comparison of gender classifications methods.," *Pattern Recognition Letters*, vol. 29, pp. 1544–1556, 2008b.

[11] B. Poggio, R. Brunelli, and T. Poggio, "Hyberbf networks for gender classification," *Proc.DARPA Image Understanding Workshop*, 1995.

[12] G. Shakhnarovich, P. A. Viola, and B. Moghaddam, "A unified learning framework for real time face detection and classification," pp. 14–21, 2002.

[13] B. Wu, H. Ai, and C. Huang, "Lut-based adaboost for gender classification," pp. 104–110, Springer-Verlag, 2003.

[14] P. J. Phillips, H. Moon, P. Rauss, and S. A. Rizvi, "The feret evaluation methodology for face-recognition algorithms," pp. 137–143, 1997.

[15] Z. E. Mayo, M., "Improving face gender classification by adding deliberately misaligned faces to the training data," *Proc. Int. Vision & Computing NZ, IVCNZ08*, pp. 1–5, 2008.

[16] L. T. Vinh, N. D. Thang, and Y.-K. Lee, "An improved maximum relevance and minimum redundancy feature selection algorithm based on normalized mutual information," pp. 395 –398, 2010.

[17] Z. Sun, G. Bebis, and R. Miller, "Object detection using feature subset selection," *Pattern Recognition*, vol. 37, no. 11, pp. 2165 – 2176, 2004.

[18] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on Neural Networks*, vol. 5, no. 4, pp. 537–550, 1994.

[19] N. Kwak and C.-H. Choi, "Improved mutual information feature selector for neural networks in supervised learning," vol. 2, pp. 1313–1318, 1999.

[20] C. A. Perez, L. A. Cament, and L. E. Castillo, "Methodological improvement on local gabor face recognition based on feature selection and enhanced borda count," *Pattern Recognition*, vol. 44, no. 4, pp. 951 – 963, 2011.

[21] M. O. B.-G. I. Priness, I., "Evaluation of gen-expression clustering via mutual information distance measure," *BMC Bio-Informatics*, vol. 111, pp. 1–12, 2007.

[22] J. Huang, Y. Cai, and X. Xu, "A filter approach to feature selection based on mutual information," vol. 1, pp. 84–89, 2006.

[23] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[24] C. M. Bishop, *Neural-network-for Pattern recognition.* Clarendon Press- Oxford. Department od Computer Science and aplications mathematics, Aston University, 1995.

[25] P. Somol, P. Pudil, and J. Kittler, "Fast branch & bound algorithms for optimal feature selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 7, pp. 900–912, 2004.

[26] M. Turk and A. Pentland, "Face recognition using eigenfaces," pp. 586 –591, jun 1991.

[27] A. M. Martinez, A. M. Mart'inez, and A. C. Kak, "Pca versus lda," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 228–233, 2001.

[28] J. Bekios-Calfa, J. Buenaposada, and L. Baumela, "Revisiting linear discriminant techniques in gender recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, pp. 858 –864, april 2011.

[29] Z. Sun, G. Bebis, X. Yuan, and S. J. Louis, "Genetic feature subset selection for gender classification: a comparison study," pp. 165–170, 2002.

[30] E. O. A. Hyvärinen, J. Karhunen, *Independent Component Analysis*. 2001.

[31] T. M. Cover and J. A. Thomas, *Elements of information Theory*. Wiley Series in Telecommunications, 1991.

[32] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," pp. 523–528, 2003.

[33] A. Jain and J. Huang, "Integrating independent components and support vector machines for gender classification," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 3, pp. 558 – 561 Vol.3, aug. 2004.

[34] X. Wang, P. Tino, M. Fardal, S. Raychaudhury, and A. Babul, "Fast parzen window density estimator," in *Neural Networks, 2009. IJCNN 2009. International Joint Conference on*, pp. 3267 –3274, june 2009.

[35] P. A. Estevez, M. Tesmer, C. A. Perez, and J. M. Zurada, "Normalized mutual information feature selection," *IEEE Transactions on Neural Networks*, vol. 20, no. 2, pp. 189–201, 2009.

[36] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," vol. 1, 2001.

[37] M. Sundaram, K. Ramar, N. Arumugam, and G. Prabin, "Histogram based contrast enhancement for mammogram images," in *Signal Processing, Communication, Computing and Networking Technologies (ICSCCN), 2011 International Conference on*, pp. 842 –846, july 2011.

[38] M.-H. Yang and B. Moghaddam, "Support vector machines for visual gender classification," vol. 1, pp. 1115–1118, 2000.

[39] M.-H. Yang and B. Moghaddam, "Gender classification using support vector machines," vol. 2, pp. 471–474, 2000.

[40] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[41] V. V. Vezhnevets, A., "Modest adaboost- teaching adabost to generalize better," *Graphicon 2005*, 2005.

[42] T. E. Bradstreet, "The analysis of means: A graphical method for comparing means, rates, and proportions," *Journal of the American Statistical Association*, vol. 101, no. 474, pp. 848–849, 2006.