Mejoras en el Entrenamiento de Esquemas de Detección de Sonrisas Basados en AdaBoost

Fernando Merchán
Facultad de Ingeniería Eléctrica
Universidad Tecnológica de Panamá
fernando.merchan@utp.ac.pa

Sebastián Galeano
Facultad de Ingeniería Eléctrica
Universidad Tecnológica de Panamá
sebastian.galeano l @utp.ac.pa

Héctor Poveda
Facultad de Ingeniería Eléctrica
Universidad Tecnológica de Panamá
hector.poveda@utp.ac.pa

Resumen- *El presente artículo aborda aspectos del entrenamiento de la máquina de aprendizaje AdaBoost con modelos de reconocimiento de objetos basados en características de apariencia tales como: Patrones Binarios Locales (LBP), Histograma de Gradientes Orientados (HOG) y características tipo Haar para la detección de sonrisas. En este contexto realizamos un estudio del impacto de varios parámetros de entrenamiento de los modelos. Proponemos un nuevo enfoque con respecto a la selección de muestras positivas utilizadas en el periodo de aprendizaje. A diferencia de otros trabajos que utilizan como muestras positivas rostros sonrientes completos, proponemos utilizar únicamente la sección del rostro correspondiente a la boca sonriente. Las pruebas realizadas muestran que nuestro enfoque ofrece hasta un 40% de disminución en el tiempo de entrenamiento y hasta un 20% de disminución en el tiempo de detección con respecto al enfoque convencional, conservando una precisión de detección comparable. Además, se estudió la influencia de la normalización del tamaño de las imágenes de entrenamiento y prueba en ambos enfoques de entrenamiento. También se estudió el impacto del tamaño de las ventanas de análisis en el rendimiento de los métodos de detección para el caso de entrenamiento usando bocas sonrientes como muestras positivas.*

Palabras Clave— *detección de sonrisas, AdaBoost, características tipo Haar, patrones binarios locales, histogramas de gradientes orientados*

Abstract- *This paper addresses training aspects of the Adaboost learning machine with object recognition models based on appearance features such: Local Binary Patterns, (LBP), Histogram of Oriented Gradients (HOG) and Haar features for smile detection. In this context, we study the impact of several training parameters in the performance of the models. We propose a new approach with respect to the selection of positive training samples. Unlike other studies that use complete smiling faces as positive samples, we propose to use only smiling mouths. The results show that our approach provides as far as a 40% reduction in training*

time and a 20% reduction for the detection time with respect to the conventional approach, achieving a very close accuracy. We also study the impact of scaling image size in training and test images in both training approaches. We also tested the impact of the size of the analysis windows when using smiling mouths as positive samples in the performance of the approaches.

Palabras Clave— *smile detection, AdaBoost, Haar features, linear binary patterns, Histogram of oriented gradients*

Tipo de Articulo: *original*
Fecha de Recepción: *15 de septiembre de 2014*
Fecha de Aceptación: *18 de noviembre de 2014*

1. Introducción

En las últimas décadas, los adelantos tecnológicos han permitido una mayor capacidad de procesamiento en toda clase de dispositivos tales como computadoras portátiles, tabletas, teléfonos celulares, sistemas embebidos, desarrollando para los mismos una gran variedad de aplicaciones.

Con el propósito de facilitar la operación o manejabilidad, se ha incrementado el interés por el desarrollo de interfaces agradables para el usuario. Estas interfaces deben ser muy intuitivas, similares a la interacción entre los seres humanos. Actualmente aumenta el interés por interfaces basadas en la detección de gestos manuales y corporales, rasgos y gestos faciales, reconocimiento de voz y el seguimiento de actividad ocular [1, 2].

Estas nuevas maneras de interactuar con la máquina, incluyen aplicaciones en la domótica para el control de televisores, equipos estéreo, control de acceso [3]; también en la interacción con robots de asistencia social, asistencia a discapacitados y atención a personas mayores; entre otras [4].

En la literatura encontramos muchos trabajos abordando la detección de rostros, el reconocimiento de rostros y la detección de expresiones faciales [5]. En los últimos años ha habido un gran interés en la detección de expresiones faciales del estado anímico o físico de las personas tales como fatiga, sueño, alegría o enojo. Por ejemplo, se han realizado trabajos para la detección de indicios de sueño en conductores de automóviles basadas en la apertura de los ojos [6]. Actualmente, algunas cámaras fotográficas comerciales cuentan con detectores de sonrisas embebidos que indican cuando los participantes de la foto sonríen [7].

Para todas estas aplicaciones es de sumo interés que los requerimientos computacionales sean lo más reducidos posibles (por ejemplo, memoria y tiempo de cálculo) a fin de que puedan ejecutarse no únicamente en tiempo real, sino también en paralelo con otras aplicaciones.

La sonrisa es una expresión facial muy común en la vida diaria de una persona. La detección de la sonrisa es de interés para determinar el grado de satisfacción del público de un evento o de un contenido multimedia, en la educación a distancia, en videoconferencias o en videojuegos.

Hay una cantidad importante de trabajos que abordan el reconocimiento de expresiones faciales y la detección de sonrisas. En [8], Shinohara y Otsu propusieron el uso de características locales de auto-correlación combinadas con un mapa de pesos de Fischer para la representación de la cara. Tian [9] estudió la influencia de la resolución de la imagen en las diferentes etapas de un sistema de reconocimiento de expresiones faciales. En [6], Kowalik et al. desarrollan un sistema que proporciona el nivel de satisfacción de la audiencia de contenidos multimedia utilizando la detección de sonrisas. Para este sistema, se utilizó una red neuronal como clasificador de un vector de características de 16 dimensiones extraído de ocho puntos de la boca. En [10], Deniz et al. proponen el uso de las características tipo Haar y un clasificador de Viola & Jones en cascada para la detección de sonrisas. Este detector es empleado en una aplicación que permite a los usuarios interactuar con un cliente de mensajería instantánea. En este trabajo un total de 5812 imágenes se emplean

para el entrenamiento del detector (es decir, 2436 positivas y 3376 negativas). El detector alcanza una precisión de 96,1% en un conjunto de 4928 imágenes de prueba.

En [11], los patrones binarios locales (LBP) y el análisis de componentes principales (PCA) se utilizan en conjunto para la detección de sonrisas. El trabajo de Cohn y Schmidt [12] presenta un estudio sobre la diferencia en amplitud y temporización entre sonrisas espontáneas y fingidas.

En [7], Whitehille et al. advierten que la mayoría del trabajo en el reconocimiento de expresiones faciales se centra en optimizar el rendimiento de los métodos sobre bases de datos de expresiones faciales recolectados en condiciones controladas en laboratorio. Estas expresiones no son naturales, sino que en su mayoría son fingidas. Es por esta razón, que crean la base de datos GENKI, compuesta de imágenes con rostros y tomadas de la web. Sobre esta base de datos los autores prueban diferentes representaciones faciales tales como características de tipo Haar, LBP, filtros de energía de Gabor y los histogramas de gradientes orientados (HOG). Llegaron a la conclusión de que se necesita de 1.000 a 10.000 imágenes de diferentes condiciones para el entrenamiento efectivo de los modelos. También estudiaron la influencia del tipo de representación y de la máquina de aprendizaje en el rendimiento del esquema de detección.

En [13], Shan propuso usar un método basado en diferencias de píxeles como características y logra obtener una precisión alta sobre la base de datos GENKI. En este trabajo también se investigó el impacto que produce la normalización de iluminación y la pose del rostro en la precisión de la detección.

En [14], Pingping et al. proponen un método para distinguir entre sonrisas espontáneas y fingidas basadas en LBP completos discriminativos de tres planos ortogonales que viene a ser un descriptor espacio-temporal local basado en apariencia.

En [5], Cruz et al. utilizan la dinámica temporal de las emociones y expresiones faciales en video con un método de muestreo inspirado de la psicología de percepción humana.

En [15], Sun & Akansu proponen un marco de trabajo basado en el reconocimiento de expresiones faciales utilizando modelos de Markov escondidos regionales describiendo el estado de atributos faciales tales como cejas, ojos y región de la boca. El sistema propuesto es utilizado para inferir el estado mental de la persona basado en expresiones faciales espontáneas.

En este trabajo, centramos nuestra atención en aspectos del entrenamiento del algoritmo Adaboost para la detección de sonrisas utilizando tres tipos de representaciones de apariencia: las características de tipo Haar, los patrones binarios locales (LBP) y los histogramas de gradientes orientados (HOG). En primer lugar proponemos un nuevo paradigma con respecto a los ejemplos positivos para el entrenamiento. La mayoría de los trabajos en la literatura usan imágenes con caras sonrientes como ejemplos positivos y las caras no sonrientes como ejemplos negativos. A diferencia de esos trabajos, proponemos utilizar la sección del rostro de la boca sonriente como ejemplos positivos. Realizamos un estudio comparativo de los modelos obtenidos utilizando ambos esquemas de entrenamiento en términos de la precisión de detección y el tiempo de entrenamiento y detección.

Además, se estudió la influencia de la normalización del tamaño de las imágenes de entrenamiento y prueba en ambos paradigmas de entrenamiento.

Por último, se estudió el impacto del tamaño de las ventanas de análisis en el rendimiento para el caso de entrenamiento usando bocas sonrientes como muestras positivas.

En la sección 2 presentamos las características de representación de imágenes y el modelo de aprendizaje utilizados. Los aspectos de implementación y entrenamiento del esquema de detección son descritos en la sección 3. En la

sección 4 se presentan las pruebas realizadas para determinar el impacto de los diferentes aspectos de entrenamiento estudiados. Presentamos las conclusiones y las perspectivas de este trabajo en las secciones 5 y 6, respectivamente.

2. Plataformas y métodos utilizados

El desarrollo este trabajo se llevó a cabo con librerías gratuitas de código abierto, utilizando el software de visión artificial de Open CV versión 2.3.1 y algunas herramientas de programación como Octave y MATLAB. Estos algoritmos recopilados y desarrollados se ejecutaron en un ordenador con sistema operativo Windows 7 de 64bits con un procesador Intel Core i7-2700K, provisto de 8GB de memoria RAM.

A continuación presentamos una breve introducción de cada una de las características de representación de imágenes y el modelo de aprendizaje que utilizamos en este trabajo.

2.1 Características Haar

El esquema de detección de objetos propuesto por Viola & Jones usando características de tipo Haar es uno de los algoritmos basados en AdaBoost más populares y de alto rendimiento [16]. El mismo ha sido utilizado en la detección de rostros, detección de peatones [17] y la detección de coches [18].

Las características de Haar son funciones rectangulares simples de 2 dimensiones en las que se varía el tamaño y la posición de recuadros blancos y negros. En [19], Lienhart propuso las características extendidas de Haar que no sólo contienen la dirección horizontal y vertical de la funciones, sino también en un ángulo de giro de 45o (ver Fig. 1). Variando la posición y orientación de los recuadros blanco y negro podemos encontrar hasta 180,000 características diferentes en una ventana de análisis de 24x24 píxeles.

Estas características se extraen buscando la diferencia entre la suma de los píxeles dentro del recuadro negro y la suma de los píxeles bajo el rectángulo blanco.

En la Fig. 2, presentamos 5 características individuales de Haar que han sido seleccionadas por el algoritmo AdaBoost como clasificadores para la detección de una sonrisa.

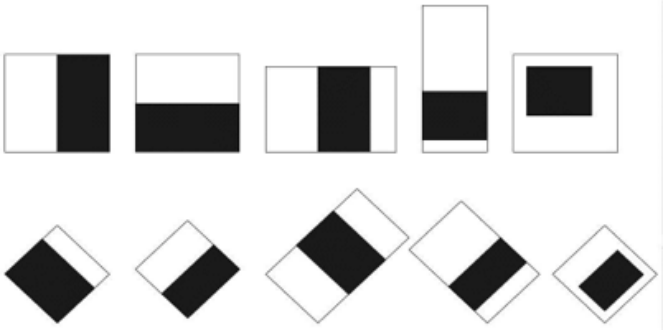

Figura 1. Ejemplo de características de Haar con diferentes posiciones y orientaciones.

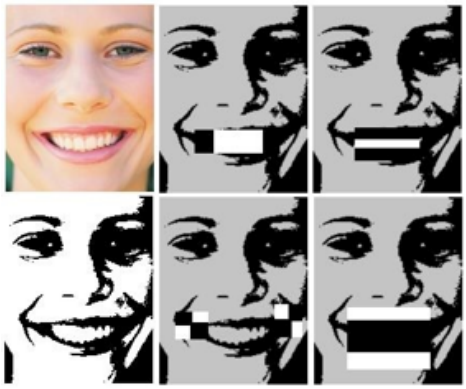

Figura 2. Características individuales de Haar que forman un clasificador fuerte para una sonrisa.

2.2 Patrones Binarios Locales (LBP)

Otro esquema muy utilizado para la extracción de características, es el presentado por Zhang et al. en [20]. En el mismo se compara la intensidad de los píxeles con sus vecinos para obtener los patrones binarios locales (LBP). Los autores lo aplican en la detección de rostro en tiempo real.

Estas características obtienen mucha más información de la estructura de la imagen que las características de Haar debido a la redundancia que genera el analizar los píxeles vecinos de cada

píxel. En algunas aplicaciones presenta mejores resultados que las características de Haar.

Este modelo define el vecindario local como un conjunto de puntos muestreados uniformemente sobre un círculo centrado en el píxel a analizar. En la figura 3 ilustramos la forma en la que se calculan los patrones LBP.

LBP ha demostrado ser altamente discriminativo por su invariancia a cambios en el nivel gris y la eficiencia computacional, lo que hacen adecuado para las exigentes tareas de reconocimiento.

Otra ventaja de LBP es que el número de características es mucho menor que las características Haar, así que el proceso de entrenamiento generalmente requiere menos tiempo. En la Fig. 4 ilustramos el cálculo de patrones binarios locales en el área de la boca.

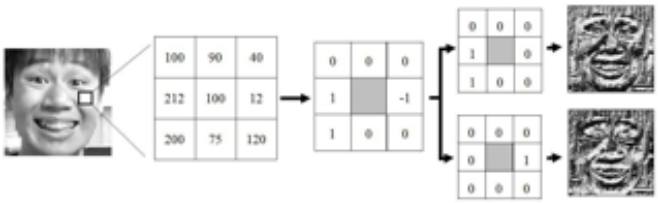

Figura 3. Cálculo de patrones binarios locales.

Figura 4. Cálculo de patrones binarios locales para rostro con sonrisa.

2.3 Histograma de Gradientes Orientados (HOG)

El último esquema utilizado para la extracción de características es el método presentado por Dalal [21]. En este modelo se representa la imagen por medio de los histogramas de gradientes de orientación (HOG). Esta representación se emplea ampliamente en la detección de peatones usando el modelo de aprendizaje y detección de Máquinas de Vectores de Soporte (SVM).

En este esquema se calculan histogramas de gradientes en celdas de *N*x*N* píxeles. Estas celdas se agrupan en bloques y un descriptor es formado por el conjunto de histogramas concatenados. Ilustramos parte de este procedimiento en la Fig. 5, donde para diferentes regiones de una imagen en escala de grises se realiza el cálculo del gradiente de la imagen. Ilustramos el resultado del procedimiento para un rostro con sonrisa abierta marcada en la Fig. 6. Esta representación captura información de los contornos de la imagen.

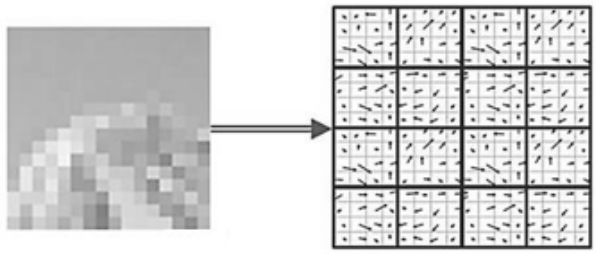

Figura 5. Cálculo de gradientes en esquema HOG.

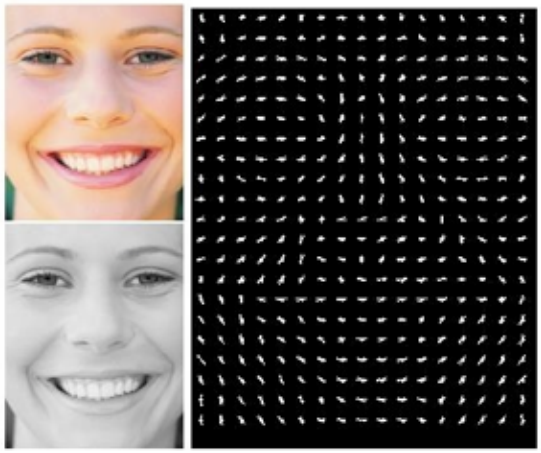

Figura 6. Cálculo de gradientes en esquema HOG para un rostro con sonrisa abierta bien marcada.

2.4 Entrenamiento y Clasificación con AdaBoost

El entrenamiento es una parte primordial en el reconocimiento de objetos, pues es el procedimiento en el que se genera un modelo estadístico que describa de manera apropiada el espacio de muestras y las variables que influyen en el proceso para el cual se entrena.

Al utilizar sistemas de detección basados en máquinas de aprendizaje, las muestras de entrenamiento son organizadas y se componen etapas, sean en árboles de decisión, clasificadores en cascadas o la estructura de una red neuronal. En este trabajo utilizamos al clasificador AdaBoost con el propósito de entrenar los modelos para cada tipo de característica de representación y reconocer las expresiones faciales en las imágenes de prueba.

El clasificador Adaboost funciona bajo la arquitectura de clasificadores en cascada. La cascada se organiza en etapas. Para cada etapa se preparan un conjunto de muestras positivas y muestras negativas, con las que se seleccionan un número de clasificadores débiles que en su conjunto forman un clasificador fuerte. Los clasificadores débiles son escogidos dentro del espacio de características de representación (Haar, LBP o HOG). Esta selección se realiza para garantizar una determinada tasa de clasificación exitosa de las muestras positivas y negativas. Detalles del algoritmo de selección de clasificadores pueden ser revisados en [16].

El análisis en cascada se efectúa sobre ventanas de análisis en la imagen a fin de determinar si la misma corresponde al espacio de muestras positivas o al espacio de muestras negativas. Si no corresponde a las muestras positivas se descarta la ventana, de lo contrario pasa a la siguiente etapa. En cada etapa se aumenta la complejidad de los clasificadores que describen las muestras positivas y las muestras negativas. Para LBP o HOG se necesita ponderar los clasificadores obtenidos para lograr eliminar patrones de histogramas repetitivos y redundantes.

El modo en el que se seleccionan los clasificadores debe seguir ciertas reglas y procedimientos que garanticen la mejor tasa de detección o de rechazo. Sin embargo como señala Ju et al. en [22] no todas las funciones e información generada por los esquemas puede ser adaptada al algoritmo AdaBoost. Para HOG por ejemplo, una característica se define con una función de N dimensiones y por lo tanto no se puede considerar una característica HOG como un clasificador débil. La característica HOG de cada celda contiene información importante sobre la manera de separar los objetos de su fondo por lo tanto, el conjunto de clasificadores débiles se crean a partir de cada celda por separado.

3. Implementación y Entrenamiento

Se consideraron dos esquemas de entrenamiento en lo que al conjunto de muestras se refiere. Un esquema que utiliza como muestras positivas caras completas y otro que utiliza únicamente la sección de la boca sonriente.

A continuación abordamos aspectos de la metodología implementada y los experimentos realizados.

3.1 Configuración de AdaBoost

Como se mencionó anteriormente, AdaBoost necesita dos grupos de imágenes: positivas y negativas. Estas muestras se organizan en etapas con el propósito de tener clasificadores fuertes en cada etapa. En las etapas de entrenamiento se utilizaron 600 ejemplos positivos. Para un esquema estos ejemplos corresponden a los rostros completos sonrientes y para otro corresponden a la sección de la boca sonriente. Se utilizaron un conjunto de 10.000 de rostros neutrales como ejemplos negativos. Para cada etapa de la cascada se utilizó el mismo conjunto de imágenes positivas y un conjunto diferente de 1.000 imágenes negativas.

La tasa de falsa alarma es un parámetro del esquema AdaBoost que corresponde a la probabilidad de falso rechazo de la hipótesis nula para una determinada prueba. La tasa de falsa alarma (FAR) se fijó entre 0,5 y 0,7. El uso de un valor en este rango permite:

1. Construir un clasificador con un mayor número de etapas.
2. Obtener un mayor número de clasificadores débiles en cada clasificador fuerte de la etapa. Por ejemplo para FAR de 0,5 se usan

50 características de Haar por etapa, mientras que para FAR de 0,7 se utilizan 90.

3. Aumentar la flexibilidad del modelo en las primeras etapas de la detección. Dado que muchas bocas neutras son similares a las sonrisas cerradas, es preferible que las primeras etapas tengan baja precisión para que la decisión de rechazo se resuelva en una etapa posterior con características más robustas y específicas.

Para el proceso de entrenamiento y formación de los modelos se probaron diferentes tamaños de ventanas de análisis para la posterior detección de las características en las imágenes. Se probaron tanto ventanas cuadradas como ventanas rectangulares. Las ventanas cuadradas condujeron a resultados con bajos rendimiento en términos de precisión de detección. Las ventanas de análisis de tamaño rectangular presentaron mejor rendimiento. En las pruebas se utilizaron ventanas con tamaños de 16x32 píxeles y 32x64 píxeles. Éstas corresponden a la proporción de las dimensiones de una boca sonriente por lo que las consideramos apropiadas para la detección de sonrisas.

3.2 Bases de datos de Entrenamiento

Utilizamos varios conjuntos de datos para la etapa de entrenamiento. Para el primer esquema de entrenamiento utilizamos rostros sonrientes de las bases de datos siguientes: UT Dallas [23], GENKI [24] y UTP como muestras positivas. Para el segundo esquema recortamos la sección de la boca sonriente del conjunto de imágenes previamente seleccionado. Como muestras negativas usamos rostros etiquetados como "sin sonreír" o "neutral" en las bases de datos siguientes: Cohn-Kanade [25], FERET [26], FEI [27].

La base de datos de rostros de la UTP es un ejercicio de recolección de imágenes de rostros realizada con estudiantes y docentes de la Universidad Tecnológica de Panamá. Ésta consistió en fotografías con iluminación controlada de 100 individuos en pose frontal, 2 fotografías por individuo; la primera sonriendo y la segunda en estado neutral.

Las especificaciones de las bases de datos utilizadas se presentan en la Tabla 1. Algunas bases de datos están a escala de grises, lo que no imposibilita su uso, debido a que la extracción de las características se realiza en un solo canal de color. Asimismo las bases de datos no poseen la misma resolución. Las sonrisas recortadas obtenidas de la base de datos de UT Dallas poseen una resolución promedio de 50x70 píxeles. Para las sonrisas obtenidas de la base de datos GENKI el tamaño varía en función de los diferentes tamaños en esta base de datos, lo que dificulta el entrenamiento de una ventana de análisis de tamaño fijo.

Presentamos en la Fig. 7, algunos ejemplos de sonrisas recortadas, de rostros positivos y rostros neutrales para ilustrar la naturaleza de las imágenes utilizadas.

Figura 7. Algunos ejemplos de sonrisas recortadas, rostros positivos con sonrisas abiertas y cerradas, y rostros en estado neutral.

Tabla 1. Especificaciones bases de datos de imágenes utilizadas.

Base de datos	# De sujetos	# De Imágenes	Espacio Color	Pose Facial	Expresión Facial	Iluminación Controlada	Resolución
FERET	1000	10000	Escala de Grises	Frontal	Neutral	Sí	256x384
Cohn Kanade	100	10000	Escala de Grises	Frontal	23 Expresiones	Sí	640x480
UT Dallas	420	840	color	Frontal	Alegre & Neutral	Sí	640x480
FEI	200	400	color	Frontal	Alegre & Neutral	Sí	360x260
GENKI	7172	7172	color	Frontal & ligera rotación	Alegre & Neutral	No	Resolución Inconstante
UTP	100	200	color	Frontal	Alegre & Neutral	Sí	480x320

3.3 Arquitectura del detector

En la literatura encontramos que la arquitectura del detector de expresiones faciales convencional en primera instancia detecta el rostro. Posteriormente, sobre éste aplica el reconocimiento de expresiones faciales, en base a un espacio de rostros completos. Finalmente encuentra un indicador con el índice de similitud más alto con respecto a la expresión facial que se desea reconocer.

La arquitectura de detección de sonrisa que proponemos consiste en las siguientes etapas:

1. *Detección de la cara:* El algoritmo de Viola-Jones para la detección de rostros se aplica para obtener la parte de las imágenes en donde se localiza una cara.
2. *Detección de sonrisas:* uno de los esquemas propuestos y entrenados (sea Haar, LBP o HOG) se aplica sobre el rostro bajo el clasificador en cascada del algoritmo AdaBoost. Esto permite detectar ventanas que presentan las características similares a las que fueron clasificadas en el modelo entrenado.
3. *Post-procesamiento:* dado que hay ocurrencias en otras partes de la cara con estructura similar a una sonrisa abierta, por ejemplo, como los ojos; se eliminan candidatos o incidencias ubicadas fuera del tercio inferior de la imagen y que presente una superficie más pequeña que el 10% del área de la cara (ver Fig. 8). Sin embargo este post-procesamiento solo se efectúa cuando el modelo detector se entrenó con bocas sonrientes; en el caso de usar caras sonrientes, la incidencia será un rostro completo.

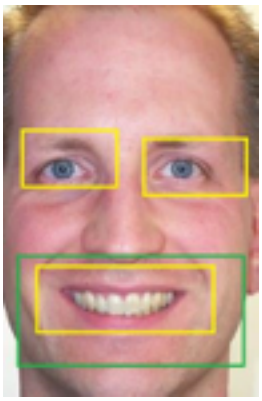

Figura 8. Etapa de post-procesamiento.

4. Pruebas y Resultados

Se realizaron pruebas en las que se estudiaron varios parámetros a fin de determinar la influencia de los mismos en el rendimiento de los esquemas de detección. Los tres parámetros estudiados son:

1. Tipo de muestras positiva para entrenamiento (es decir, modelos entrenados con caras completas sonrientes versus modelos entrenados con sonrisas recortadas.)
2. Normalización del tamaño de las imágenes de entrenamiento y prueba.
3. Tamaño de las ventanas de análisis.

Las bases de datos seleccionadas para el entrenamiento están detalladas en la sección de bases de datos, al igual que las bases de datos utilizadas para las pruebas. Para conformar 600 imágenes positivas se recolectaron 100 imágenes de la base de dato FERET, 200 de UT Dallas, 200

de FEI y 100 de UTP. Para las pruebas, hemos seleccionado imágenes de las bases de datos Cohn-Kanade, FEI, FERET y GENKI. Asimismo cabe destacar que para FERET y Cohn-Kanade hemos usado imágenes diferentes a las utilizadas en el entrenamiento. Con el respecto a la normalización del tamaño de las imágenes hemos considerado dos escenarios:

1. El tamaño de imágenes de prueba y de entrenamiento fue re-escalado y normalizado a un tamaño de 240x32. En el caso particular de las imágenes con sonrisas recortadas se hizo el ajuste proporcional en el tamaño.
2. El tamaño en las imágenes de entrenamiento y prueba no fue modificado, es decir, se trabajó con las resoluciones originales que se presentan en la Tabla 1.

El último parámetro que se tomó en cuenta es el tamaño de la ventana de análisis, el cual corresponde al tamaño mínimo en el cual se puede detectar un objeto en la imagen. Este a su vez define la resolución y tamaño de los descriptores para el esquema Haar. Las características LBP y HOG se organizan con histogramas por celdas, por esto, la ventana de análisis corresponde, para estos dos modelos, a las celdas con los tamaños de los bloques bajo los cuales se calculan los histogramas. Cuatro aspectos han sido evaluados para los diferentes escenarios:

1. Tiempo de entrenamiento: tiempo que toma al ordenador conformar el modelo.
2. Tiempo de detección promedio: tiempo que toma al ordenador procesar las detecciones e incidencias negativas en una imagen.
3. Porcentaje de verdaderos positivos: índice de imágenes que contenían sonrisas detectadas correctamente por el modelo.
4. Porcentaje de verdaderos negativos: índice de imágenes que no contenían sonrisas que no fueron detectadas como sonrisas.

Los resultados de los escenarios de influencia de la naturaleza de la base de datos y la normalización de tamaño son presentados en Tablas 2 y 3. En la Tabla 2 se resumen los resultados de las imágenes de entrenamiento y prueba en las que no se ha normalizado el tamaño, tanto para rostros sonrientes completos como para sonrisas recortadas. En la Tabla 3 se resumen los resultados para el caso con re-escalamiento de tamaño, sea para rostros sonrientes completos como para sonrisas recortadas. Las tablas están elaboradas de tal manera que los datos a la izquierda corresponden al rendimiento asociado al uso de caras completas como base de pruebas positiva, mientras que el lado derecho presentan los resultados del uso de bocas sonrientes como muestras positivas.

Los resultados de las Tablas 2 y 3, demuestran que el tiempo de entrenamiento disminuye entre un 25-40% en los diferentes esquemas cuando se usan bocas sonrientes en comparación con rostros sonrientes completos. El tiempo empleado en la detección de los modelos que utilizan bocas sonrientes son 5-25% inferiores al tiempo de los modelos que utilizan caras completas como muestras positivas.

La precisión de los modelos con boca sonriente y cara completa es comparable. Estos poseen un margen de variación en las tasas de detección de verdaderos positivos y verdaderos negativos de hasta ±5%. Esto representa en las pruebas una diferencia máxima de 10 imágenes de un total de 200 imágenes.

Por otro lado, se constata una disminución de 5-20% en tiempo de entrenamiento cuando usamos imágenes normalizadas en tamaño en rostros completos en relación con entrenamiento con rostros completos sin normalizar. Sin embargo este tiempo no incluye el tiempo de re-escalado de la imagen que corresponde a 7 ms por imagen aproximadamente.

Los índices de detección de verdaderos positivos entre modelos normalizados y no normalizados en tamaño oscilan entre ±4.5% (mejorando y empeorando el desempeño), lo que

para un análisis sobre 200 imágenes de prueba significa un total de 9 imágenes de discrepancia en rendimiento, lo cual es un margen relativamente pequeño.

Este mismo comportamiento lo observamos en el índice de verdaderos negativos, con diferencias en los índices oscilando en ±4.0%, significando un margen de 8 imágenes, lo que es una diferencia relativamente pequeña. Los modelos de las Tablas 2 y 3 han sido entrenados bajo los parámetros señalados en la Tabla 4.

Con respecto a los tres esquemas utilizados, las tasas de detección son comparables. Por tanto, los tres esquemas son apropiados y robustos para la detección de sonrisas. En tiempo de entrenamiento LBP posee hasta un 25% de reducción en comparación con Haar y hasta una mejoría de 15% frente a HOG. Los tiempos de detección en cambio no representan una diferencia considerable, un valor máximo de 5% de reducción usando LBP frente a HOG y 8% frente a Haar.

El modelo de Haar posee en las imágenes analizadas muchas incidencias fuera de la región de la boca, que se descartan gracias a la etapa de post-procesamiento. Esta problemática que no ocurre con LBP o HOG que poseen características mucho más precisas con respecto a los vecinos de los píxeles como a las regiones pequeñas de análisis. LBP y HOG son mucho más robustos en la detección en imágenes con baja resolución.

Los resultados de los escenarios en los que entrenamos sistemas con diferentes ventanas de análisis podemos observarlos en las Tablas 5 y 6.

Para estas pruebas se utilizó el entrenamiento con bocas sonrientes como ejemplos positivos y se utilizaron imágenes no normalizadas en tamaño. Se tomaron como muestras positivas 200 sonrisas recortadas de imágenes de rostros de la base de datos UT Dallas. Para cada etapa se utilizaron como muestras negativas 1,000 imágenes de rostros en estado neutral de las bases de datos Cohn-Kanade y FERET.

En la Tabla 5, presentamos los resultados de entrenamiento con ventanas de análisis de 16x32, mientras que en la Tabla 6 presentamos los resultados con las mismas imágenes usando ventanas de análisis de 32x64.

Tabla 2. Modelos entrenados no normalizados en tamaño, caras completas sonrientes versus bocas sonrientes.

Base de Datos Negativa	FERET NO NORMALIZADA EN TAMAÑO					
Base de Datos Positiva	CARAS COMPLETAS CON SONRISA			BOCAS SONRIENTES		
Característica	HAAR	LBP	HOG	HAAR	LBP	HOG
Tiempo entrenamiento (horas)	8,72	7,01	7,94	5,35	4,69	4,68
Base de Datos de Prueba:	FERET NO NORMALIZADA					
Tasa de positivos verdaderos (%)	99,62	97,61	97,68	96,78	95,98	96,63
Tasa de negativos verdaderos (%)	98,85	99,80	98,11	97,57	98,82	98,55
Tiempo de detección promedio (ms)	14,79	14,19	14,03	12,52	11,62	11,45
Base de Datos de Prueba:	FEI NO NORMALIZADA					
Tasa de positivos verdaderos (%)	97,59	88,65	96,49	95,88	94,15	98,80
Tasa de negativos verdaderos (%)	98,95	97,28	96,84	93,34	96,93	97,24
Base de Datos de Prueba:	15,22	14,26	13,41	12,02	11,84	11,94
Tasa de positivos verdaderos (%)	COHN KANADE NO NORMALIZADA					
Tasa de negativos verdaderos (%)	94,79	98,89	99,82	93,14	95,15	94,01
Tiempo de detección promedio (ms)	91,17	95,29	96,47	91,65	98,68	95,21
Base de Datos de Prueba:	14,50	14,34	13,89	12,71	11,94	11,72

Comparando los datos de las tablas, podemos observar una disminución en el tiempo de entrenamiento de los modelos de 16x32 con respecto al de 32x64 de un 5-20%. Para los modelos con ventanas de análisis de 16x32 el tiempo de detección promedio aumento en un 5-40% conrespecto a los modelos de 32x64, este deterioro es ocasionado por el hecho de que le toma al algoritmo más tiempo procesar un número mayor de ventanas pequeñas que un conjunto pequeño de ventanas grandes. Este efecto es producido por la naturaleza de los clasificadores que usamos en AdaBoost.

Los índices de detección fueron superiores en los casos en los que la imagen de prueba es de baja resolución (por ejemplo, en GENKI) y la ventana es de 16x32. Para el modelo con ventana de 32x64 se obtuvieron incidencias de falsos positivos mucho más grandes que las sonrisas que se pretendían detectar en imágenes de baja resolución como GENKI.

En general los índices más estables de positivos verdaderos y falsos verdaderos se logran con modelos de 16x32, debido a la resolución de los clasificadores débiles. Al ser mucho más pequeña, describen mejor la estructura de las sonrisas, pero estas ventajas no corresponden a una diferencia muy grande en tasas de aciertos.

Queremos destacar que para las pruebas presentadas en las Tablas 5 y 6 se utilizaron sólo 200 muestras positivas de entrenamiento lo cual constituye un número bajo en relación con lo que se presenta en la literatura. A pesar de esto los resultados de detección presentan precisiones bastante buenas (por ejemplo, de 98% frente a un 92% con 400 muestras menos de entrenamiento).

Tabla 3. Modelos entrenados normalizados en tamaño, caras completas sonrientes versus bocas sonrientes.

Base de Datos Negativa	FERET NORMALIZADA EN TAMAÑO					
Base de Datos Positiva	CARAS COMPLETAS CON SONRISA			BOCAS SONRIENTES		
Característica	HAAR	LBP	HOG	HAAR	LBP	HOG
Tiempo entrenamiento (horas)	8,36	5,95	7,57	6,58	4,83	5,99
Base de Datos de Prueba:	FERET NORMALIZADA EN TAMAÑO					
Tasa de positivos verdaderos (%)	97,90	98,75	98,05	96,36	97,02	96,56
Tasa de negativos verdaderos (%)	97,81	93,87	95,79	95,68	94,66	99,05
Tiempo de detección promedio (ms)	13,48	14,78	13,33	11,14	11,80	11,73
Base de Datos de Prueba:	FEI NORMALIZADA					
Tasa de positivos verdaderos (%)	90,38	97,75	99,96	98,04	95,04	97,62
Tasa de negativos verdaderos (%)	95,79	97,35	95,65	99,65	97,55	98,54
Base de Datos de Prueba:	14,41	13,59	13,45	11,50	11,49	11,02
Tasa de positivos verdaderos (%)	COHN KANADE NORMALIZADA					
Tasa de negativos verdaderos (%)	93,88	96,35	97,90	98,56	98,46	95,60
Tiempo de detección promedio (ms)	93,78	96,43	96,42	89,57	98,73	97,14
Base de Datos de Prueba:	14,46	13,70	14,20	11,42	11,62	11,01

Tabla 4. Parámetros de entrenamiento para las Tablas 2 y 3.

Característica	HAAR	LBP	HOG
Muestras Positivas	600	600	600
Muestras negativas por etapa	1000	1000	1000
Etapas	20	20	20
Ventana análisis	[16 32]	[16 32]	[16 32]
False Alarm Rate	0.6	0.6	0.6
True Positive Rate	0.995	0.995	0.995

Tabla 6. Modelos entrenados con base de datos de bocas sonrientes y ventanas de análisis de 32x64.

Base de datos de Entrenamiento	Cohn-Kanade			FERET		
Característica	HAAR	MB-LBP	HOG	HAAR	MB-LBP	HOG
Tiempo de entrenamiento (h)	6.95	3.43	4.74	5.51	2.36	3.82
BaseDatos de Prueba	FEI			FEI		
Tasa de positivos verdaderos (%)	94.09%	85.38%	92.15%	99.00%	99.00%	96.50%
Tasa de negativos verdaderos (%)	86.40%	90.10%	93.85%	94.50%	87.90%	86.00%
Tiempo de detección promedio (ms)	13.94	12.70	15.75	13.26	12.29	15.69
BaseDatos de Prueba	GENKI			GENKI		
Tasa de positivos verdaderos (%)	83.00%	84.40%	83.50%	84.70%	88.20%	85.20%
Tasa de negativos verdaderos (%)	77.30%	78.1%	80.50%	73.70%	76.20%	77.90%
Tiempo de detección promedio (ms)	10.20	10.73	11.44	11.48	10.45	11.03
BaseDatos de Prueba	FERET			Cohn-Kanade		
Tasa de positivos verdaderos (%)	95.80%	93.00%	90.20%	81.30%	82.90%	86.6%
Tasa de negativos verdaderos (%)	88.00%	87.50%	82.00%	90.8%	93.40%	91.10%
Tiempo de detección promedio (ms)	11.48	11.15	12.47	13.41	12.53	13.97

5. Conclusiones

En este trabajo se estudiaron varios aspectos referentes al entrenamiento de esquemas de detección de sonrisas basados en el algoritmo AdaBoost utilizando tres tipos de representaciones: características de tipo Haar, LBP y HOG.

Se propuso un nuevo paradigma para el entrenamiento de los modelos de detección que consiste en el uso de bocas sonrientes como muestras positivas en vez del rostro completo con sonrisa. Los resultados muestran que usar las bocas sonrientes reduce el tiempo de entrenamiento hasta en un 40% y reduce hasta en un 25% el tiempo de detección. Al modificar el espacio de las muestras de entrenamiento logramos optimizar las distribuciones probabilísticas de estas imágenes en la detección. Esto permite reducir el número de muestras necesarias para el aprendizaje.

Asimismo se estudió la influencia de normalizar por re-escalamiento el tamaño de las imágenes de entrenamiento y prueba. Los resultados arrojaron que la precisión de los datos no varía de forma significativa. Esto indica que los esquemas de representación presentan cierta robustez a variaciones en la resolución.

Igualmente se estudió el impacto del tamaño de las ventanas de análisis. Los menores tiempos de entrenamiento se obtuvieron con la ventana de análisis de menor tamaño, de 16x32 píxeles. Sin embargo la ventana de 32x64 presenta tiempos de detección inferiores.

Las pruebas demostraron que el desempeño de las características LBP es superior a Haar y éste a su vez superior a HOG. Sin embargo LBP posee mucha más información redundante y un tiempo de detección mucho más alto que los demás.

6. Trabajos Futuros

Actualmente, muchos sistemas de análisis de expresión facial intentan determinar expresiones faciales directamente en categorías emocionales básicas como lo hemos como fue el objetivo de este trabajo. Obtuvimos resultados aceptables y comparables, mejorando el tiempo de detección y el tiempo de entrenamiento. Sin embargo, estos tipos de esquemas no son naturalmente apropiados para manejar las acciones faciales causadas por las actividades diarias. Posiblemente el análisis del sistema de codificación de acciones faciales (FACS) puede proporcionar una solución a este reto, ya que permiten clasificar las acciones faciales antes de cualquier intento de interpretación frente a la manera en que se producen dinámicamente; de otro modo, este procedimiento no es algo que se pueda implementar con facilidad con estos esquemas que hemos puesto a prueba.

Se deben además implementar e integrar sistemas de análisis de expresión facial autónomos, métodos de extracción de características dependientes de aprendizaje dinámico, inicialización automática de entrenamiento y detección de la expresión de mayor índice de similitud.

Más investigaciones deben efectuarse con el fin de combinar otros sistemas como un lector de ritmo cardiaco, frecuencia respiratoria, integración de voz, tono y lenguaje, como también gestos corporales.

7. Agradecimiento

Este trabajo es parcialmente financiado por la Secretaría Nacional de Ciencia, Tecnología e Innovación de Panamá (SENACYT).

Referencias Bibliográficas

[1] S. Bodiroza; G. Doisy; V.V. Hafner, "Position-invariant, real-time gesture recognition based on dynamic time warping," Human-Robot Interaction (HRI), 2013 8th ACM/IEEE International Conference on , vol., no., pp.87,88, 3-6 March 2013.

[2] R. Sato; Y. Takeuchi, "Coordinating turn-taking and talking in multi-party conversations by controlling robot's eye-gaze," Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on , vol., no., pp.280,285, 25-29 Aug. 2014.

[3] L. Sang-Heon; S. Myoung-Kyu; K. Dong-Ju; K. Byungmin; K. Hyunduk, "Smart TV interaction system using face and hand gesture recognition," Consumer Electronics (ICCE), 2013 IEEE International Conference on , vol., no., pp.173,174, 11-14 Jan. 2013.

[4] Z. Xinshuang; A.M. Naguib; L. Sukhan, "Kinect based calling gesture recognition for taking order service of elderly care robot," Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on , vol., no., pp.525,530, 25-29 Aug. 2014.

[5] A, Cruz.; B, Bhanu ; N, Thakoor, "Vision and Attention Theory Based Sampling for Continuous Facial Emotion Recognition," Affective Computing, IEEE Transactions on, (aceptado para publicación)

[6] U. Kowalik, T. Aoki, and H. Yasuda, "Broaference—a next generationmultimedia terminal providing direct feedback on audience's satisfaction level," in Proc. IFIP TC13 Int. Conf. Human-Comput. Interact. (INTERACT), 2005, pp. 974–977.

[7] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan, "Towards practical smile detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 11, pp. 2106–2111, Nov. 2009.

[8] Y. Shinohara and N. Otsu, "Facial expression recognition using fisher weight maps," in Proc. IEEE Int. Conf. Autom. Face Gesture Recog., 2004, pp. 499–504.

[9] Y.Tian, "Evaluation of Face Resolution for Expression Analysis" IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'04), Washington, DC, USA, 2004.

[10] O. Deniz, M. Castrillon, J. Lorenzo, L. Anton, and G. Bueno, "Smile detection for user interfaces," in Proc. Int. Symp. Adv. Vis. Comput., 2008, pp. 602–611.

[11] D. Freire, M. Castrillon, and O. Deniz, "Novel approach for smile detection combining LBP and PCA," in Proc. Int. Conf. Comput. Aided Syst. Theory (EUROCAST), 2009.

[12] J. F. Cohn and K. L. Schmidt, "The timing of facial motion in posed and spontaneous smiles," Int. J.Wavelets, Multiresolution Inf. Process., vol. 2, pp. 1–12, 2004.

[13] C. Shan,."Smile Detection by Boosting Píxel Differences". IEEE Trans. on Image Processing, Vol. 21, No. 1, January 2012.

[14] W. Pingping; L. Hong; Z. Xuewu, "Spontaneous versus posed smile recognition using discriminative local spatial-temporal descriptors," Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on , vol., no., pp.1240,1244, 4-9 May 2014.

[15] S. Yanjia; A.N. Akansu, "Automatic inference of mental states from spontaneous facial expressions," Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on , vol., no., pp.719,723, 4-9 May 2014.

[16] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in. Proc. of the 2001 IEEE Conference on Computer Vision and Pattern Recognition , vol.1, 2001, pp.I-511,I-518.

[17] R. Miyamoto, H. Sugano and H. Saito, "Pedestrian recognition in far-infrared images by combining boosting-based detection and skeleton-based stochastic tracking", in Advances in Image and Video Technology Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2006, 483-494.

[18] R. N. Hota, K. Jonna and P.R. Krishna, "On-road vehicle detection by cascade classifiers", in Proceedings of the 3rd Bangalore Annual Compute Conference, Bangalore, India, January 22-23, 2010.

[19] R. Lienhart, J. Maydt, "An extended set of Haar-like features for rapid object detection", in Proc. International Conference on Image Processing, vol.1, 2002, pp.900-903.

[20] L. Zhang, R. Chu, S. Xiang and S. Liao, "Face detection based on Multi-Block LBP representation" , in Advance in Biometrics: Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2007.

[21] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection", in Proc of the Conf. on Computer Vision and Pattern Recognition, (2005, pp. 886-893.

[22] Y. Ju, H. Zhang and Y. Xue, "Research of Feature Selection and Comparison in AdaBoost based Object Detection System", Journal of Computational Information Systems, pp. 8947–8954, 2013.

[23] M. Minear and D. C. Park,. A lifespan database of adult facial stimuli. Behavior Research Methods, Instruments, & Computers, 36, 630–633, 2004. http://agingmind.utdallas.edu/facedb, The UT Dallas Face Database.UT Dallas Subset.

[24] The MPLab GENKI Database, Available: http://mplab.ucsd.edu, GENKI-4K Subset.
[25] T. Kanade, J. F. Cohn, J. F., and Y.Tian., "Comprehensive database for facial expression analysis". Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition (FG'00),Grenoble,France, 2000. pp.46-53, http://www.pitt.edu/ ~emotion/ck-spread.html. Cohn-Kanade Dataset.
[26] P. J. Phillips, H. Wechsler, J. Huang and P. J. Rauss: The FERET database and evaluation procedure for face-recognition algorithms. Image Vision Comput., pp.295-306.1998. Available : http://www.itl.nist.gov/iad/humanid/feret/feret_master.html.
[27] The FEI Database, FEI Subset Available: http://fei.edu.br/~cet/facedatabase.html.