SEP TecNM

TECNOLÓGICO NACIONAL DE MÉXICO
INSTITUTO TECNOLÓGICO DE ACAPULCO

TEMA:

IDENTIFICACIÓN AUTOMÁTICA DE ESTADOS EMOCIONALES BÁSICOS A TRAVÉS DEL ANÁLISIS FACIAL.

OPCIÓN I:
TESIS PROFESIONAL

QUE PARA OBTENER EL TÍTULO DE:
MAESTRO EN SISTEMAS COMPUTACIONALES

PRESENTA:
ING. MARIO JIMÉNEZ VÁZQUEZ

DIRECTOR DE TESIS:
DR. JOSÉ ANTONIO MONTERO VALVERDE

CO- DIRECTOR DE TESIS:
DRA. MIRIAM MARTÍNEZ ARROYO

Noviembre de 2018

Dedicatoria

Esta tesis esta dedicada:

A mi padre Filemón†, a pesar de nuestra distancia física, siento que estás conmigo siempre y aunque nos faltaron muchas cosas por vivir juntos, sé que este momento hubiera sido tan especial para tí como es para mí.

A mi madre Lorenza, por ser el pilar más importante y por demostrarme siempre su cariño y apoyo incondicional, quien con su amor, paciencia y esfuerzo me ha permitido llegar a cumplir hoy un sueño más, gracias por inculcar en mí el ejemplo de esfuerzo y valentía para no temer a las adversidades porque Dios está siempre conmigo.

A mis hermanos por su cariño y apoyo incondicional, durante todo este proceso, por estar conmigo en todo momento gracias. A toda mi familia porque con sus oraciones, consejos y palabras de aliento hicieron de mí una mejor persona y de una u otra forma me acompañan en todos mis sueños y metas.

Finalmente quiero dedicar esta tesis a mi esposa Olivia, y a mis hijas Ivana y Diana por apoyarme cuando más las necesito, por extender su mano en los momentos difíciles y por el amor brindado cada día, de verdad mil gracias, siempre las llevo en mi corazón.

Agradecimientos

Quiero expresar mi gratitud a Dios, quien con su bendición llena siempre mi vida y me fortalece para seguir adelante, y a toda mi familia por estar siempre presentes en los momentos más difíciles.

Mi agradecimiento al Consejo Nacional de Ciencia y Tecnología (CONACyT), por el apoyo otorgado para llevar a cabo este trabajo, al Instituto Tecnológico de Acapulco, y a mis profesores en especial a la Dra. Miriam Martínez Arroyo y al Dr. Eduardo de la Cruz Gámez quienes con la enseñanza de sus valiosos conocimientos hicieron que pueda crecer día a día como profesional, gracias a cada uno de ustedes por su paciencia, dedicación, apoyo incondicional y amistad.

Finalmente quiero expresar mi más grande y sincero agradecimiento al Dr. José Antonio Montero Valverde, principal colaborador durante todo este proceso, quién con su dirección, conocimiento, enseñanza y colaboración permitió el desarrollo de este trabajo.

Contenido

Índice de Figuras

Índice de Tablas.

Capítulo 1 Introducción

1.1 Antecedentes

Las emociones que los humanos expresan a través del rostro juegan un papel relevante en la vida social. Son señales visualmente observables, conversacionales e interactivas que determinan nuestro foco de atención y regulan nuestra interacción con el entorno y personas vecinas [46]. Asimismo, sabemos que actualmente las computadoras se están convirtiendo en parte de nuestras vidas. Invertimos una cantidad razonable de nuestro tiempo interactuando con dispositivos computacionales de uno u otro tipo (celulares, tabletas, iPhone, videojuegos, etc.). Por el momento, estos dispositivos son, generalmente, indiferentes al estado emocional de las personas. Sin embargo, es del conocimiento común que, para conducir una comunicación humano-humano efectiva debemos tener la habilidad de detectar las señales emocionales de los demás. Por lo tanto, una interacción humano-máquina que no toma en cuenta los estados afectivos de los usuarios pierde una gran parte de información disponible la cual se considera relevante en esta tarea.

El análisis de la expresión facial nos proporciona información relacionada con varias funciones, por ejemplo: regular la conversación, enfatizar un discurso, mostrar conocimiento, regular y mostrar el estado emocional.

Recientemente, la investigación relacionada con los estados afectivos ha sido ampliamente estudiada y existe una creencia creciente de que proveer a las computadoras con la capacidad de entender los estados emocionales de las personas es una tarea importante [61], [71]. Se cree que, con el fin de conseguir progresos en el futuro de las interacciones humano-máquina

es necesario que éstas puedan reconocer el estado emocional de los usuarios. Esto se da por entendido debido a la importancia que tienen las emociones en nuestras vidas [9]. La computación afectiva es una rama de investigación que estudia el enlace entre los humanos como entes emocionalmente afectivos y las máquinas como dispositivos con deficiencia emocional [13].

Existen muchas áreas de aplicación que saldrían beneficiadas si las máquinas tuvieran la habilidad de reconocer emociones en usuarios humanos. Algunas de estas aplicaciones son, entre otras: i) tutores inteligentes que adaptan la enseñanza si observan que el estudiante está cansado o confundido, ii) interfaces que no interrumpen si detectan que el usuario está muy atento en su tarea, iii) videojuegos que adaptan su dificultad en base a la atención y respuesta del usuario, iv) sistemas de salud que avisan a los médicos si detectan que el paciente está sintiendo dolor, v) tecnologías que diagnostican estados como la depresión, y vi) dispositivos que monitorean a choferes detectando señales de cansancio o aburrimiento.

Estos sistemas serán de utilidad a medida que reconozcan de manera confiable el estado anímico de las personas con las que interactúan. Los humanos nos expresamos de una forma compleja y multimodal. Las personas se comunican e interpretan a través de expresiones faciales, movimientos de las manos, señales auditivas, postura, movimientos de la cabeza y movimientos de los ojos. A través de estas vías de comunicación se transmite información importante que las personas utilizan para inferir el estado emocional de los demás [1]. De estas modalidades, las expresiones faciales son las que han recibido más atención por parte de psicólogos e investigadores de la computación [89]. Esto no es casualidad, pues el rostro es la parte social visible del cuerpo humano. A través de las expresiones faciales se pueden mostrar emociones [23], se comunican intenciones y ayudan a regular la interacción social

[75]. Aunque no se considera parte del rostro, los gestos con la cabeza también juegan un papel importante en la comunicación visible y ha sido foco de interés en la investigación para la detección de afecto [68].

De acuerdo con la teoría neuro-cultural de Ekman [21], el ser humano muestra seis emociones básicas -felicidad, enojo, disgusto, sorpresa, tristeza y miedo- las cuales se asocian con la activación autónoma de patrones y expresiones faciales. Asimismo, la habilidad para distinguir estas emociones a través del análisis facial tiende a ser universal entre los humanos. Se han realizado varios de trabajos relacionados con la identificación automática de las seis emociones básicas, extendiéndose estas con el uso de estándares (FACS, FAPS)[1] a la identificación de más de cien emociones individuales [22]. Aunque diseñar un sistema que solamente reconozca seis emociones no resulte de gran utilidad en la mayoría de los escenarios, si puede ser de gran apoyo cuando lo que se busca es la detección específica de algunas emociones para cierta aplicación.

Por ejemplo, resulta de gran utilidad que los sistemas tutores inteligentes integren una herramienta para conocer el estado emocional de los usuarios y determinar si se encuentra en una condición que le permita aprovechar una sesión de aprendizaje. En estos casos, más que diseñar una herramienta computacional compleja que permita adaptarse a diferentes escenarios, resulta de mayor utilidad contar con una herramienta sencilla que pueda integrarse en dispositivos que no requieran muchos recursos computacionales. En este sentido, el diseño de identificadores de emociones específicas (cansancio, aburrimiento, etc.) que operen en condiciones ambientales normales resultan muy útiles.

[1] FACS (Facial Action Code System, (Ekman and Friesen, 1977)); FAPs (Facial Animation Parameters, (Cowie et al, 2008)).

En este trabajo se presenta el diseño de una arquitectura sencilla que permite el reconocimiento confiable de estados emocionales básicos a través del análisis de expresiones faciales basadas en imágenes obtenidas en condiciones ambientales reales.

1.2 Objetivo General

Diseñar una herramienta computacional que identifique de manera confiable las emociones básicas expresadas por humanos.

1.3 Objetivos Específicos

- Crear una base de imágenes de personas en ambientes no controlados y mostrando emociones básicas.

- Desarrollar una herramienta que detecte de manera confiable el rostro de una persona en una imagen estática obtenida en ambientes reales.

- Desarrollar un clasificador que ofrezca un rendimiento aceptable, aun cuando no se tengan suficientes datos para el entrenamiento.

1.4 Hipótesis

El diseño e implementación de una herramienta computacional que permita la identificación confiable de las emociones humanas básicas es fundamental para el diseño de interfaces naturales humano-máquina más complejas.

1.5 Metodología

La figura 1.1 muestra la metodología utilizada en este trabajo con el fin de reconocer de manera automática cuatro emociones humanas básicas. Como se observa, la metodología consta de cinco etapas. En la etapa 1 se obtiene el rostro de una persona utilizando la cámara de una computadora, la imagen se toma bajo condiciones ambientales reales. En la etapa 2 se identifica el rostro de una persona en la imagen tomada con anterioridad, para esto se aplica el algoritmo propuesto por Viola y Jones [83]. Una vez que el rostro es detectado en la imagen se procede a la alineación de la misma aplicando la técnica de los Promedios de Filtros Sintéticos Exactos (ASEF), esto se realiza en la etapa 3. Para realizar la extracción de las características que representan las diferentes emociones faciales, se utiliza la técnica de Histograma de Gradientes Orientados (HOG) [15], esta tarea se lleva a cabo en la etapa 4. El aprendizaje del modelo basado en las Máquinas de Vectores Soporte (MVS) utilizando las características seleccionadas en el paso anterior se realiza en la etapa 5 [11], en esta etapa se tienen que considerar el cincuenta por ciento de las imágenes para el entrenamiento y el otro cincuenta por ciento de las imágenes almacenadas en la base de datos para la evaluación del modelo.

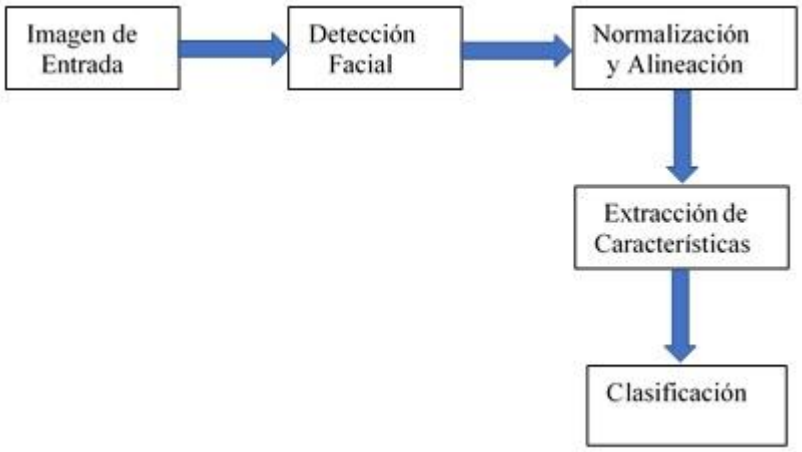


Figura 1.1 Etapas para el reconocimiento de emociones básicas a través del análisis de expresiones faciales.

1.6 Organización de la Tesis

La organización de este trabajo se describe a continuación: El capítulo 2 muestra una recopilación de los trabajos más relevantes que se relacionan con el presentado. El Capítulo 3 describe los fundamentos del marco teórico utilizado. El Capítulo 4 describe la metodología utilizada y desarrollo de la misma. La evaluación del modelo, pruebas y resultados, se `presentan en el Capítulo 5. Finalmente, el capítulo 6 muestra las conclusiones y trabajo futuro.

Capítulo 2 Trabajos Relacionados

2.1. Introducción

El estudio de las expresiones faciales se ha analizado y estudiado desde sus inicios en el Siglo XVII. En el año de 1649, John Bulwer escribe una nota detallada sobre las diversas expresiones y movimiento de los músculos de la cabeza en su libro "Pathomyotomia". En 1667, Le Brun académico y pintor francés ofreció una conferencia relacionada con su trabajo sobre las expresiones faciales en la Real Academia de Pintura que más tarde fue reproducida en forma de libro en el año de 1734 [55]. Posteriormente en 1872, Charles Darwin escribió un tratado donde estableció los principios generales de expresión en los seres humanos y los animales [18]. Esto creó mucha controversia en el momento de su publicación debido a reclamo polémico de la universalidad de las emociones y sus orígenes evolutivos. Asimismo, el trabajo de Darwin muestra el agrupamiento de varios tipos de expresiones faciales en categorías similares.

2.2 Sistemas Estándares para las Expresiones Faciales: FACS y FAPS

Los diferentes comportamientos faciales y los movimientos que se realizan en el rostro se pueden describir mediante parámetros en base a las acciones musculares que se realizan. Este conjunto de parámetros se puede utilizar para representar las diferentes expresiones faciales. Considerando que las expresiones faciales se han venido estudiando desde hace muchos años, ha habido dos trabajos importantes y exitosos en la creación de estos conjuntos de parámetros que han derivado en el establecimiento de dos estándares.

1.- Sistema de Codificación de Acción Facial (FACS), desarrollado por Ekman y Friesen en 1977 [22, 25].

2.- Parámetros de Animación Facial (FAPs), Sistemas de Parametrización Facial que fueron implementados Pandzic y Forcheimer [58].

A continuación, se hace una breve descripción de estos estándares.

2.2.1.-El Sistema de Codificación de Acción Facial (FACS)

Antes de que apareciera el sistema de codificación de acción facial, los investigadores ponían a un grupo de humanos a observar los rostros de las personas y después de un cierto periodo de tiempo debían entregar su análisis. Desde luego estas observaciones visuales no podían ser consideradas como un resultado exacto debido a que los datos proporcionados por los observadores dependían de varios factores no controlados tales como: la distancia, la iluminación o el ángulo de observación, por lo que estos resultados no podían ser muy confiables y precisos. Debido a estos elementos, que influían en los resultados de las observaciones, Ekman y otros investigadores pusieron en evidencias tales resultados argumentando que los observadores podían estar influenciados por el contexto y que además las observaciones que realizaban podían no ser las mismas en todas las culturas [19].

Debido a los factores descritos, las limitaciones que tenían los observadores podían ser superadas mediante la representación de las expresiones faciales y los comportamientos en términos de un conjunto fijo de parámetros faciales. Con un marco de este tipo en su lugar, solamente estos parámetros individuales tienen que ser observados sin tener en cuenta el

comportamiento facial como un todo. A pesar de que desde principios de la década de 1920 los investigadores estaban tratando de medir las expresiones faciales y desarrollar un sistema basado en parámetros, ningún consenso había surgido y los esfuerzos fueron muy dispares. Para resolver estos problemas, Ekman y Friesen en el año de 1977, desarrollaron el sistema integral de FACS, que desde entonces ha convertido en un estándar universal en estos estudios [22].

La Codificación de Acción Facial es un enfoque basado en el análisis de los músculos faciales. Se trata de identificar los diversos músculos faciales que individualmente o en grupos causan cambios en las expresiones del rostro. Estos cambios en la cara y los derivados en los músculos que los causan reciben el nombre de *unidades de acción* (UA). El Sistema Codificación de Acción Facial (FACS) se compone de varias de estas *unidades de acción* [25].

2.2.2 Parámetros de Animación Facial (FAP's)

Los sistemas basados en los parámetros faciales son una parte del estándar MPEG-4, Codificación Híbrido Natural Sintética (SNHC)[2] [81], (trabajo basado en el análisis de una secuencia de imágenes (video) realizado en el año de 1998 para el análisis y estudio de los parámetros faciales).

[2] SNHC (Synthetic Natural Hybrid Coding, (Video, M. P. E. G. (1998). SNHC,"Text of ISO/IEC FDIS 14 496-3: Audio,". *Atlantic City MPEG Mtg*.)).

Antes de la década de 1990, la comunidad de investigación y animación por computadora se enfrentó a problemas similares que los investigadores de reconocimiento de expresión del rostro tuvieron antes de que aparecieran los Sistemas de Codificación de Acción Facial (FACS). No había ninguna norma unificadora y cada sistema de animación desarrollado tenía su propio conjunto definido de parámetros. Como se ha señalado por Pandzic y Forcheimer [58], los esfuerzos de los investigadores de la animación y la graficación estaban más centrados en los movimientos faciales y en los parámetros utilizados, en lugar de los esfuerzos para elegir el mejor conjunto de parámetros [58]. Este enfoque hacía inservibles los sistemas a través de dominios. Para hacer frente a estos problemas y proporcionar un control sobre el conjunto de parámetros utilizados para tener una expresión facial estandarizada, el grupo de expertos en imágenes en movimiento (MPEG)[3] presentó las especificaciones de animación facial (FA) en el estándar MPEG-4.

La versión 1 del estándar MPEG-4 (junto con la especificación FA) se convirtió en el estándar universal a partir de 1999.

En los años posteriores, los investigadores dedicados al reconocimiento de expresión de la cara, empezaron a utilizar el estándar MPEG-4, estos estándares son métricas para modelar las expresiones faciales [12]. El estándar MPEG-4 es compatible con la animación facial, proporcionando parámetros de animación facial (FAPs).

[3] MPEG (Moving Pictures Experts Group (Grupo de expertos en imágenes en movimiento fundado por Hiroshi Yasuda,1988)).

Cowie y sus colegas indican la relación entre los MPEG-4, FAP y FACS, centrándose principalmente en la síntesis de la expresión facial y la animación que define los parámetros de animación facial (FAPs) los cuales están fuertemente relacionados con las Unidades de Acción (UA), el núcleo de los FACS [12].

Desde la década de 1990, la investigación sobre el reconocimiento automático de la expresión facial se ha convertido en un tema muy activo. Se encuentran disponibles estudios completos y ampliamente citados en [60], así como en [27], que realizan un estudio en profundidad del trabajo publicado desde años atrás. Esto ha llevado al mismo tiempo al desarrollo de la interacción humano-computadora, debido a estos estudios la computación afectiva comenzó a popularizarse.

2.3 Computación Afectiva

Recientemente la computación afectiva ha sido ampliamente estudiada por diferentes investigadores y su popularidad creció más con la publicación del libro "Affective Computing" de Rosalind Picard, y se tenía la creencia de poder diseñar computadoras que tuvieran la capacidad de leer los estados afectivos de sus usuarios [67], [62], [71]. Se cree que para tener un progreso en la Interfaz Humano-Computadora (HCI), es necesario reconocer el estado afectivo de los usuarios, esto se basa en la importancia de la emoción en la vida diaria de los humanos de acuerdo con los conceptos presentados [9].

La computación afectiva trata de cerrar la brecha que existe entre las computadoras emocionalmente deficientes y las personas emocionalmente expresivas [13]. La detección

automática del afecto ha atraído un gran interés de diversos campos y grupos de investigación, incluidos la psicología, las ciencias cognitivas, la lingüística, la visión por computadora, el análisis del habla y el aprendizaje automático. Se cree que con la inclusión del afecto en estos campos impulsarán en gran medida su desarrollo.

Este campo de las ciencias computacionales ha crecido y se ha diversificado en las últimas décadas, abarcando la detección automática del afecto a través de las expresiones faciales, concluyendo con el diseño y desarrollo de interfaces emocionalmente inteligentes. La computación afectiva es un campo demasiado amplio para describir en detalle en este trabajo, pero se intenta proporcionar una visión general del mismo con énfasis en la detección del afecto mediante el análisis de las expresiones faciales.

La tecnología computacional se ha desarrollado muy rápidamente debido a las necesidades observadas y soluciones que se dan para satisfacer a los usuarios, y cada vez se desarrollan nuevas técnicas y algoritmos para este fin. En este sentido, se ha visto la inclusión de nuevas estrategias en las diferentes etapas de la visión computacional, tal como el empleo de las Máquinas de Soporte Vectorial (MVS´s) desarrolladas por Vapnik en el año de 1995 como un mecanismo eficiente de clasificación [11]. Así mismo, también han surgido técnicas para extraer de forma eficiente características de imágenes digitales. Este es el caso de los histogramas de gradientes orientados (HOG), los cuales son técnicas de descripción basados en vectores de gradientes [15].

2.4 Técnicas Utilizadas en el Reconocimiento Facial

En el reconocimiento de las emociones humanas se han utilizado varios enfoques. La técnica de Análisis de Componentes Principales (PCA) combinadas con las Redes Neuronales Artificiales (ANN), fueron utilizadas por Filko y Martinovic, para reconocer siete expresiones faciales (Neutro, Miedo, Felicidad, Tristeza, Enojo, Disgusto y Sorpresa) [31]. Así mismo, se utilizó en la etapa de preprocesamiento en estas imágenes el filtrado de Canny para resaltar los bordes en el rostro de las personas y facilitar el proceso de extraer características mediante las PCA. Posteriormente se utilizó una red neuronal artificial (ANN) para clasificar las características de cada emoción. Este proceso se aplicó a un conjunto de imágenes estáticas almacenadas en la base de datos FEEDTUM[4] (figura 2.1) de la universidad de Munich [31]. Los autores reportan una eficiencia del 46.00% al 80.00%, en el reconocimiento de las emociones faciales utilizando las técnicas mencionadas. Las imágenes utilizadas fueron extraídas de la base de datos FEEDTUM.

En este mismo contexto, Gosavi y Koth realizaron la identificación de las siete emociones humanas básicas (Neutro, Miedo, Felicidad, Tristeza, Enojo, Disgusto y Sorpresa) considerando el uso del filtrado de Sobel durante la etapa de preprocesamiento con el fin de resaltar las orillas en la imagen [38]. Posteriormente, aplicaron la técnica de Análisis de Componentes Principales para extraer y representar la información de características correspondientes a los siete gestos observados.

[4] FEEDTUM (Facial Expressions and Emotions Database from the Technical University of Munich).

En este trabajo fueron utilizadas 213 imágenes almacenadas en la base de datos JAFFE[5] (figuras 2.2, 2.3, 2.4). Para las pruebas de entrenamiento se utilizaron imágenes de diez personas cada una mostrando siete emociones diferentes (una imagen para cada expresión facial, 70 imágenes). Para el reconocimiento de estas imágenes se les realizó un ajuste de tamaño de 256 x 256 pixeles. Los autores mencionan que al evaluar su modelo éste arroja un 67.14 % de reconocimiento de las expresiones faciales con las imágenes utilizadas [38].

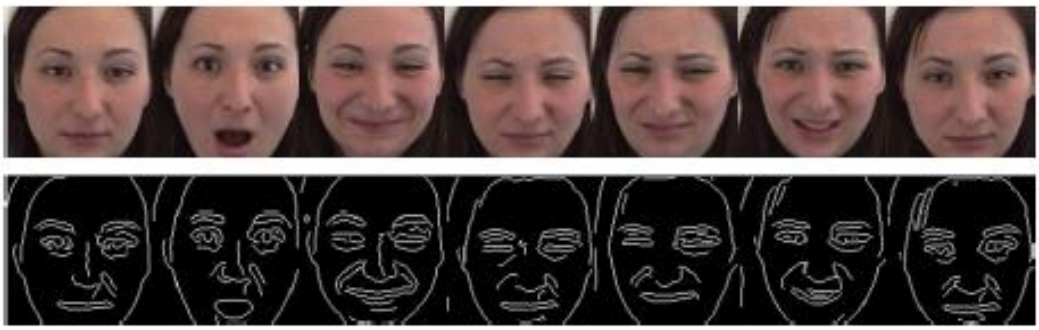


Figura 2.1. Siete emociones universales de un sujeto en la base de datos FEEDTUM [31].

En figura 2.1 de la base de datos FEEDTUM [31], las imágenes superiores muestran los rostros expresando diferentes emociones. En las imágenes inferiores, se muestra el resultado después de aplicar el filtrado utilizando la técnica de Canny, a las imágenes superiores. Como se observa en la imagen, este filtrado resalta los bordes y el contorno del rostro mostrando las expresiones faciales detectadas en el mismo.

Otros investigadores también continuaron con las disertaciones sobre las expresiones faciales, para llevar a cabo estos estudios utilizaron la base de datos japonesa JAFFE [76].

[5] JAFFE (Japanese Female Facial Expression).

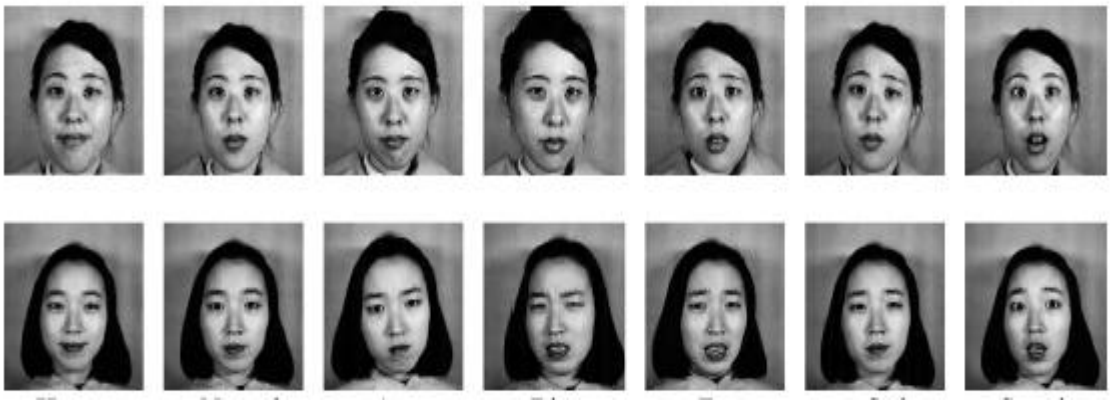


Figura 2.2. Muestras de siete expresiones faciales diferentes de la Base de Datos JAFFE [76].

En la figura 2.2 se muestran imágenes de rostros femeninos almacenados en la base de datos JAFFE, que fueron utilizadas para detectar siete diferentes emociones en mujeres japonesas.

Para el reconocimiento automático de emociones humanas realizado por Zhang y otros [88], los autores utilizaron la información proporcionada por un video obtenido a través de una webcam. En este trabajo se integraron dos técnicas de extracción de características: las basadas en la técnica de Análisis de Componentes Principales (PCA) y la técnica de Patrones Locales Binarios (Local Binary Pattern, LBP). Al combinar estas técnicas, los autores reportan los siguientes resultados. Al utilizar la combinación de técnicas PCA y MVS el porcentaje de reconocimiento alcanzado fue 91.25 %. Para el sistema integrado PCA+LBP+SVM el reconocimiento fue del 93.75 %.

El diseño de sistemas automatizados para reconocer los estados emocionales a través del análisis facial sigue siendo un tema de interés. En este sentido, Thuseethan y Kuhanesan [79], diseñaron un sistema que detecta las siete emociones humanas básicas descritas en los

trabajos anteriores. Al igual que estos trabajos, utilizaron la técnica de Análisis de Componentes Principales (PCA) para extraer y representar las características faciales. Y, al igual que los trabajos mencionados utilizaron la base de datos JAFFE (figuras 2.2, 2.3, 2.4).

En este caso, utilizaron 213 imágenes en escala de grises para detectar los estados emocionales de: felicidad, tristeza, sorpresa, enojado, neutral, disgusto y miedo. El sistema detecta la cara bajo condiciones controladas para proporcionar un rendimiento óptimo, por ejemplo: sin gafas, sin vello facial y la cabeza no debe hacer un movimiento rígido. Por lo tanto, la imagen de entrada debe satisfacer las condiciones anteriores para realizar la detección facial. Es decir, bajo condiciones controladas.

Neutral Miedo Feliz Enojado Disgusto Tristeza Sorpresa

Figura 2.3. Imágenes con diferentes emociones utilizadas para entrenamiento de la base de datos JAFFE [76].

Durante el preprocesamiento el sistema convierte la imagen a escala de grises eliminando el color, las imágenes utilizadas son de un tamaño de 480 x 480 píxeles para su posterior procesamiento. En el trabajo desarrollado por Thuseethan y Kuhanesan [79], utilizando la técnica de PCA, se reportó una tasa de acierto en la tarea de clasificación de 91.16%.

En el año 2008, con la finalidad de hacer una comparación entre las técnicas de Análisis Linear Discriminativo (2D Linear Discriminant Analysis-2D-LDA), Máquina de Vectores de Soporte (MVS), Análisis de Componentes Principales (PCA) y Red neuronal artificial de Función de Base Radial (Radial Basis Function Network-RBFN), Shih y otros [76], presentaron un trabajo para el reconocimiento de las siete emociones humanas básicas utilizando la base de datos JAFFE. En las imágenes faciales se utilizaron dos estrategias para la extracción de características. Para la primera, se dividió la base de datos en diez segmentos de diferentes expresiones en el cual la combinación de 2D-LDA con MVS presentó tasas de reconocimiento de 94.13%. Para la segunda estrategia se aplicó solo un segmento a una imagen de una expresión facial, en el cual el resultado fue de 95.10%.

En una investigación realizada por Rao, Saroj, Maity y Koolagudi [37], fue presentado un video en el cual se aplicó la técnica de Redes Neuronales Artificiales (ANNs), para el reconocimiento de cinco emociones (Felicidad, Tristeza, Enojo, Miedo y Neutro). En este trabajo fueron consideradas tres regiones de interés para la tarea de reconocimiento facial: el ojo izquierdo, el ojo derecho y la boca, para este proceso fue diseñada una red neuronal artificial (ANN) para reconocer cada región y cada emoción. En este trabajo, el porcentaje de reconocimiento reportado por los autores fue del 87%.

Otra investigación realizada en el año de 2014 por Pérez-Gaspar y otros [65], el reconocimiento de cuatro emociones (Enojo, Felicidad, Tristeza y Neutro) fue llevado a cabo con una Red Neuronal Artificial (ANN), optimizada con un Algoritmo Genético sobre la base de datos JAFFE. Para dicho trabajo las regiones de los ojos y boca fueron extraídas y una ANN para el modelado de error de reconocimiento fue integrada para mejorar el

desempeño de la ANN principal. La tasa de reconocimiento reportada en este trabajo fue del 85 %.

Feliz Enojado Neutro Triste

Figura 2.4. Muestra de Imágenes de la base de datos emocional JAFFE [76].

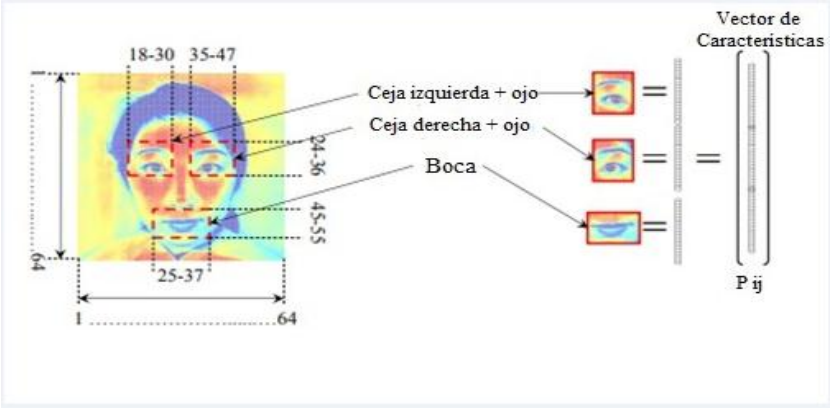


Figura 2.5. Regiones consideradas para la extracción de características [65].

El reconocimiento automatizado de las expresiones faciales es un tema que ha generado diferentes opiniones por los investigadores ocasionando con ello algunas discusiones en el tema, algunas de estas se relacionan con las diferencias que existen entre las razas y etnias de los pueblos con diferentes culturas, porque se creía que las expresiones faciales de las personas de un estado o país podían ser diferentes a las expresiones de un individuo de otra región, por lo que se tuvo que someter a un nuevo estudio. Este caso dio origen a que se

diseñara una base de datos de personas mexicanas (BD-MX) (figura 2.6) para incluirla en los estudios de las expresiones faciales junto a la base de datos JAFFE [86], [87].

En la base de datos mexicana se incluyeron las imágenes de nueve usuarios mexicanos del Este y Suroeste del país, (tres hombres y seis mujeres) a quienes se les solicitó que expresaran las emociones de Enojo, Felicidad, Tristeza y Neutro [64] [65]. Tres muestras de cada emoción fueron capturadas para realizar pruebas con la base de datos MX. Estas imágenes fueron tomadas con un fondo blanco y en condiciones de iluminación estándar [48]. Esto condujo a tener almacenadas 108 imágenes en la base de datos (9 usuarios x 3 muestras x 4 emociones), aplicando las técnicas de PCA+ANN+GA para el reconocimiento de emociones faciales, dio como resultado 83.74% de reconocimiento a las dos bases de datos.

Figura 2.6. Muestra de imágenes preprocesadas de las Bases de Datos MX y JAFFE [48].

En otra investigación Lyons y sus colegas [49], hicieron uso del filtro de Gabor en diferentes escalas y orientaciones, y lo aplicaron en varios puntos referenciales para cada imagen convolucionada para construir el vector de características que representa cada una de las emociones. Después de realizar ese proceso, aplicaron el Análisis de Componentes

Principales (PCA) para reducir la dimensionalidad de la función de vectores, y utilizaron el Análisis Discriminativo Linear (LDA) para identificar las siete expresiones faciales diferentes. La tasa de reconocimiento reportada fue de 92%.

Otros investigadores que se enfocaron en el reconocimiento facial fueron Zhang y otros [88] quienes en su trabajo utilizaron los coeficientes en cascada de Gabor y las posiciones geométricas con el fin de construir el vector de características para cada imagen y aplicaron una red neuronal artificial del tipo perceptrón de dos capas para clasificar siete expresiones faciales diferentes. La tasa de reconocimiento reportada fue del 90.1%.

La Tabla 2.1 muestra un concentrado de trabajos relacionados con el reconocimiento de expresiones faciales que representan emociones humanas básicas. La información que se muestra consiste básicamente en las técnicas utilizadas durante la extracción y representación de las características faciales, clasificador empleado y el porcentaje de reconocimiento alcanzado.

Tabla 2.1 Técnicas utilizadas para el reconocimiento facial

Autor	**Base de Datos**	**Técnicas Utilizadas**	**Estados Emocionales**	**% de Reconoc.**
A.P Gosavi, S.R. Khot,	JAFFE	PCA	Neutro, miedo, felicidad, tristeza, enojo, disgusto, sorpresa	67.14%
D. Filko, G. Martinovic,	FEEDTUM	PCA-ANN	Neutro, miedo, felicidad, tristeza, enojo, disgusto, sorpresa	46.00% - 80.00%

Y. Luo, Wu, C., Y. Zhang,	Video	MVS	Neutro, miedo, felicidad, tristeza, enojo, disgusto, sorpresa	71.50%
Y. Luo, Wu, C., Y. Zhang,	Video	PCA+MVS	Neutro, miedo, felicidad, tristeza, enojo, disgusto, sorpresa	91.25%
Y. Luo, Wu, C., Y. Zhang,	Video	PCA+LBP+ MVS	Neutro, miedo, felicidad, tristeza, enojo, disgusto, sorpresa	93.75%
S. Thuseethan, S. kuhanesan,	JAFFE	PCA	Felicidad, tristeza, ira, sorpresa, disgusto, miedo.	91.16%
F.Y. Shih, C.-F.Chuang, P.S.P. Wang,	JAFFE	2D-LDA+MVS	Neutro, miedo, felicidad, tristeza, enojo, disgusto y sorpresa	95.10%
F.Y. Shih, C.-F.Chuang, P.S.P. Wang,	JAFFE	2D-LDA+MVS	Neutro, miedo, felicidad, tristeza, enojo, disgusto y sorpresa	94.13%
K.S. Rao, V.K. Saroj, S. Maity, S.G. Koolagudi,	Video	ANN	Felicidad, tristeza, enojo, miedo y neutro	73.00% - 87.00%
L.A. Perez-Gaspar, S. O. Caballero-Morales, F. Trujillo-Romero,	JAFFE	ANN-GA	Enojo, felicidad, tristeza y neutro	85.00%
L.A. Perez-Gaspar, S. O. Caballero-Morales, F. Trujillo-Romero,	JAFFE + BD-MX	PCA	Enojo, Felicidad, Neutro, Tristeza	78.25 %
L.A. Perez-Gaspar, S. O. Caballero-Morales, F. Trujillo-Romero,	JAFFE + BD-MX	PCA-ANN	Enojo, Felicidad, Neutro, Tristeza	77.90 %

L.A. Perez-Gaspar, S. O. Caballero-M. F. Trujillo-Romero,	JAFFE + BD-MX	PCA+ANN+ AG	Enojo, Felicidad, Neutro, Tristeza	83.74%
M. Lyons, J. Budynek, S. Akamatsu	JAFFE	PCA+LDA	Neutro, miedo, felicidad, tristeza, enojo, disgusto y sorpresa	92.00%
Z. Zhang, M. Lyons, M. Schuster, S. Akamatsu	JAFFE	Filtros de Gabor + ANN-2 capas	Neutro, miedo, felicidad, tristeza, enojo, disgusto y sorpresa	90.10%

Capítulo 3 Marco Teórico

3.1 Introducción

La identificación o clasificación de emociones humanas de forma automática es una tarea desafiante, debido a la variedad de característica propias de cada persona para expresar sus emociones, aunado a esto intervienen otros aspectos, tales como: cambios significativos en la iluminación, posición del rostro, color y escalamiento. Por lo tanto, en este trabajo se propone la integración de dos técnicas que han mostrado su eficiencia en trabajos similares: los descriptores basados en Histogramas de Gradientes Orientados (HOG) [15], y las Máquinas de Vectores Soporte (MVS) [11].

Los descriptores HOG son una técnica que ha demostrado su robustez ante pequeñas variaciones en el contorno de una imagen, posición y dirección, y a cambios significativos en la iluminación y al color. Asimismo, las MVS, han demostrado su confiabilidad en la tarea de clasificación aún con una cantidad limitada de datos de entrenamiento. Los elementos mencionados son aspectos que caracterizan la tarea de identificar las emociones humanas a partir del análisis facial.

En este trabajo se intenta clasificar de forma confiable las emociones humanas básicas o elementales ante condiciones reales de iluminación y con una cantidad limitada de datos de entrenamiento.

Un desafío en los trabajos relacionados con visión computacional consiste en encontrar un descriptor capaz de funcionar de forma apropiada ante variaciones de iluminación y en un amplio rango de poses. Este es un aspecto que se ha manejado en trabajos anteriores para la

identificación automática de objetos y de personas, y donde los autores han reportado resultados satisfactorios aplicando HOG´s y MVS´s [72], [57], [2].

A continuación, se hace una descripción de las técnicas principales que integran la arquitectura utilizada en este trabajo. La explicación se va a realizar apoyándonos en uno de los trabajos que la han utilizado [15].

3.2 Histograma de Gradientes Orientados

La idea principal detrás de los Histogramas de Gradientes Orientados consiste en que la apariencia y forma de un objeto en una imagen pueden ser caracterizados como una distribución de gradientes de las intensidades, es decir, por las orientaciones del contorno u orillas. Esta caracterización se realiza al dividir la imagen en pequeñas regiones conectadas (celdas) a partir de las cuales se obtienen los gradientes. A su vez, estas celdas se agrupan en regiones más grandes llamadas bloques, los cuales representan valores de intensidad que al ser normalizados representan a todas las celdas. El efecto que resulta al normalizar los valores de los bloques es volver al descriptor invariante ante los cambios de iluminación.

La técnica de los histogramas de gradientes orientados fue presentada por primera vez en 1986 por Robert K. McConnell sin utilizar el término de Histogramas de Gradientes Orientados en una solicitud de patente. El uso de los HOG's se generalizó en el año 2005 por Dalal y Triggs [15] - investigadores del Instituto Nacional Francés para la Investigación en Ciencias de la Computación y Automatización (INRIA), quienes presentaron un trabajo

complementario sobre los descriptores HOG en la Conferencia sobre Visión por Computadora y Reconocimiento de Patrones (CVPR).

3.2.1. ¿Qué es un Descriptor de Características?

Un descriptor de características es una representación de una imagen o una sección de la misma que la simplifica al extraer información útil y descartar información irrelevante. Típicamente, un descriptor de características convierte una imagen (matriz 2D) a un vector de longitud ***n*** (conjunto de ***n*** características**)**. Por ejemplo, si tenemos una imagen de entrada de tamaño 64 x 128 pixeles, el vector de características calculado tiene una longitud igual a 3780 datos.

A continuación, se describe como se obtiene este valor:

En una imagen de tamaño 64 x 128, se determinan celdas con un tamaño de 8 x 8 píxeles, posteriormente se determina un bloque de mayor tamaño para normalizar, en este caso será de 16 x 16 pixeles, (cuatro celdas: dos celdas verticales y dos celdas horizontales). La orientación de los gradientes calculados en cada celda se va a numerar en intervalos de 9 bins, los cuales se concatenan para formar un descriptor. Ya que cada bloque está formado por cuatro celdas, se tienen 36 bins por bloque.

La imagen se va a dividir en bloques de 16 x 16 pixeles, con un 50% de traslape, esto genera una imagen con 105 bloques en total (7 x 15). Cada celda muestra una representación de los gradientes (magnitud y dirección) a través de histogramas. Considerando que la característica

consiste en el gradiente representado a través de un histograma por cada celda, la longitud del vector de características calculado en esta imagen es de 105 x 4 x 9 = 3780 datos.

En el descriptor de características, la distribución (histogramas) de las direcciones de los gradientes se utilizan como características. Los gradientes (derivadas en ***x*** e ***y***) de una imagen son útiles porque su magnitud es grande alrededor de los bordes y esquinas (regiones de cambios abruptos de intensidad) y sabemos que los bordes y esquinas contienen más información sobre la forma del objeto que las regiones planas[6] del mismo.

3.2.2 Obtención de un Histograma de Gradientes Orientados

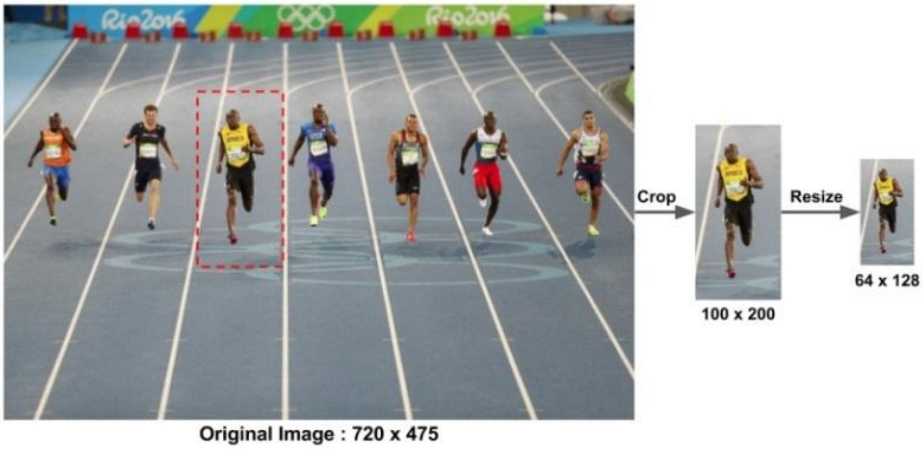
Como se mencionó con anterioridad el descriptor de características HOG utilizado para la detección de personas en una imagen se calcula en una sección de **n x n** pixeles. Por supuesto, una imagen puede ser de cualquier tamaño. Normalmente, las secciones en múltiples escalas se analizan en muchas ubicaciones de la imagen. La única restricción es que las secciones que se analizan deben tener una relación de tamaño de 1: 2. Por ejemplo, pueden ser 100 × 200 o 1000 × 2000 y en imágenes 32 x 64 o 128 × 256 pero no 101 × 205.

Para ilustrar este aspecto, se muestra una imagen de tamaño 720 × 475 pixeles (figura 3.1). Se selecciona una sección de la misma de tamaño 100 × 200 (rectángulo de líneas punteadas en la figura 3.1) para calcular el descriptor de características HOG.

Esta sección se recorta de la imagen y se redimensiona a un tamaño de 64 × 128 pixeles.

Una vez que se tiene esta dimensión de la imagen se procede como se indicó en la subsección 3.2.1 a obtener el descriptor HOG.

6 Regiones con intensidad homogénea.



a) b) c)

Figura 3.1. a) Imagen original, b) imagen recortada, c) imagen de 64 x 128 pixeles [52].

3.2.3 Cálculo de Imágenes de Gradientes

Para calcular un descriptor HOG, primero se deben calcular los gradientes horizontales y verticales de la imagen. La finalidad de esto es resaltar algunos aspectos de la imagen tales como las orillas o bordes. Esto se puede realizar utilizando filtros, máscaras o kernels, tales como los mostrados en la figura 3.2, los cuales acentúan los bordes en los ejes ***x*** e ***y*** de una imagen.

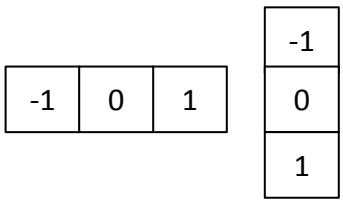


Figura 3.2. Máscaras aplicadas a una imagen para acentuar bordes.

A continuación, podemos encontrar la magnitud y dirección del gradiente usando la siguiente fórmula:

$$G = \sqrt{Gx(m,n)^2 + Gy(m,n)^2} \quad 3.1$$

$$\theta = arc\tan\left(\frac{Gy(m,n)}{Gx(m,n)}\right) \quad 3.2$$

Donde:

m = [1, 2, 3, … ,m], filas de la máscara.

n = [1, 2, 3, … ,n], columnas de la máscara.

G = magnitud del gradiente.

Θ = orientación (ángulo) del gradiente en grados.

Con las fórmulas descritas se calcula la dirección y la magnitud del gradiente en un hiperplano de dos dimensiones, considerando que todos los descriptores tienen una orientación medida en ángulos de 0 a 180 grados**.** En cada píxel de una imagen, el gradiente tiene una magnitud y una dirección. Para las imágenes en color, se evalúan los gradientes de los tres canales. La magnitud del gradiente en un píxel es el máximo de la magnitud de los gradientes de los tres canales, y el ángulo es el correspondiente al del gradiente máximo.

Cálculo del Histograma de Gradientes en Celdas de 8 × 8 Pixeles. En este paso, la imagen se divide en celdas de 8 × 8 pixeles y se calcula un histograma de gradientes para cada celda (figura 3.3).

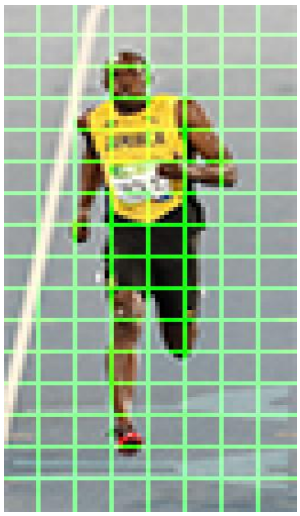


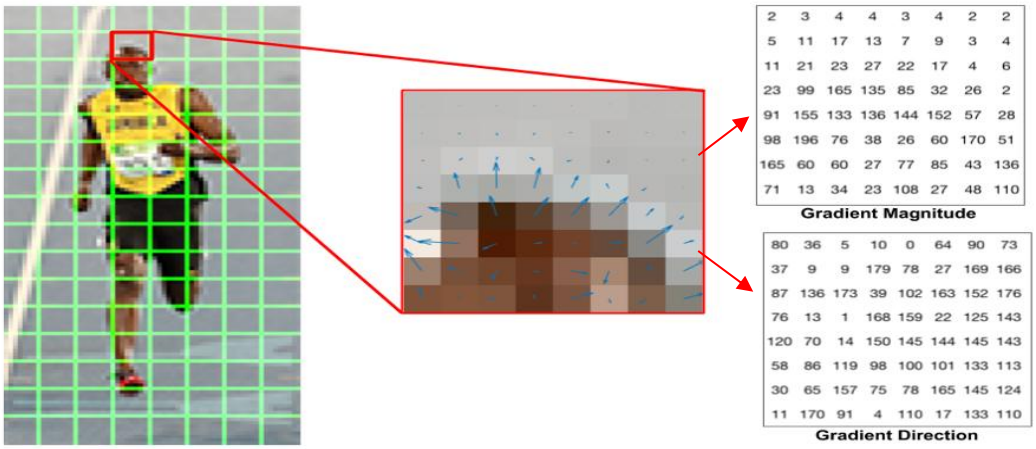
Figura 3.3. Imagen dividida en celdas de 8 x 8 pixeles [52].

Una de las razones para usar un descriptor de características para describir una sección de una imagen es que proporciona una representación compacta del contenido de la imagen con todos los atributos contenidos en la misma. Una sección de imagen de 8 × 8 pixeles, contiene 8 x 8 x 3 (canales RGB) = 192 valores de píxeles. El gradiente de esta sección está representado por 2 valores (magnitud y dirección) por píxel, por lo tanto, cada celda tiene un total de 8 x 8 x 2 = 128 valores de este tipo. No solo es la representación más compacta, el cálculo de un histograma sobre una sección de la imagen hace que esta representación sea más robusta al ruido. Los gradientes individuales pueden tener ruido, pero un histograma en una sección de 8 × 8 pixeles hace que la representación sea más compacta y mucho menos sensible al ruido.

Pero, ¿por qué un tamaño de 8 × 8? ¿Por qué no 32 × 32? Es una opción de diseño informada por la escala de características que estamos buscando. HOG se utilizó inicialmente para la detección de peatones. Las celdas de 8 × 8 pixeles en una foto de un peatón convertida a escala de 64 × 128 pixeles son lo suficientemente grandes como para capturar características

interesantes (por ejemplo, la cara, la parte superior de la cabeza, etc.). El histograma consiste esencialmente en una representación escalada en intervalos de 9 bins correspondientes a los ángulos 0, 20, 40, 60 ... 180 grados.

La figura 3.4 muestra un corte de la imagen de tamaño 8 × 8 pixeles, la ampliación muestra la dirección del gradiente al aplicarle los filtros de acentuamiento mostrados en la figura 3.2, los valores de gradientes y magnitud de los mismos se muestran en las imágenes de la derecha.



a) b) c)

Figura 3.4. a) Imagen dividida en celdas, b) Recuadro de la imagen, c) Magnitud y dirección del gradiente mostrado en números correspondiente a los grados [52].

3.2.4 Orientación

Cada píxel dentro de la celda indica un voto ponderado para el histograma basado en orientación, considerando los valores encontrados en el cálculo del gradiente. Las celdas

pueden ser analizadas de forma rectangular o radial, esto depende del tipo de kernels utilizados para acentuar los bordes. Para obtener la orientación de la región donde se observa el máximo gradiente se aplica la fórmula 3.2, asimismo y con el fin de eliminar ruido se puede aplicar un filtrado adaptable que elimina el ruido pero no suaviza los bordes. Dalal y Triggs [71] en su investigación encontraron que particionar la imagen en tamaños de 8 x 8 píxeles, así como la división de los ángulos en 9 bins (figura 3.5) proporcionó un mayor rendimiento en sus experimentos para la detección de personas en una imagen. En cuanto al peso del voto, la contribución del píxel puede ser la magnitud del gradiente en sí o alguna función de la magnitud. En las pruebas, la magnitud del gradiente en sí misma generalmente produce los mejores resultados. Otras opciones para el peso del voto podrían incluir la raíz cuadrada o el cuadrado de la magnitud del gradiente, o alguna versión recortada de la magnitud.

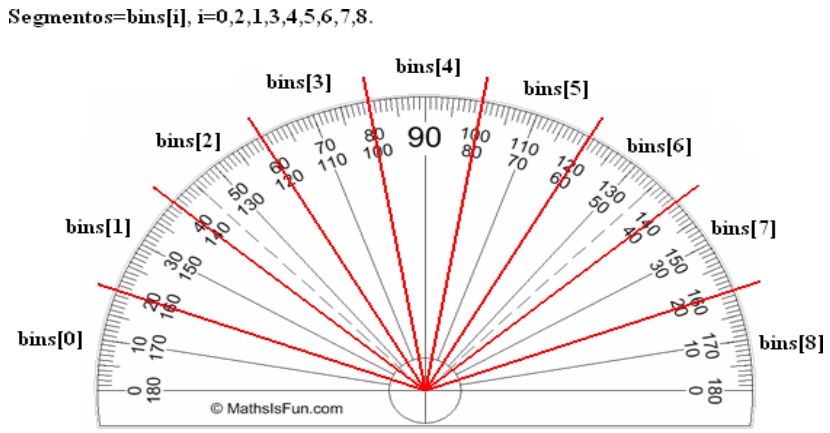


Figura 3.5. Intervalos de orientación de ángulos de 0 a 180^0 [41].

3.3 Cómo se Calcula la Magnitud y Dirección del Gradiente

Para crear un histograma de gradientes para la sección de la imagen con celdas de 8 × 8 pixeles, se deben calcular previamente, aplicando las expresiones algebraicas descritas por

las Ecs. 3.1 y 3.2, los valores de la magnitud y dirección del gradiente. El histograma del gradiente en este caso está compuesto por 9 intervalos (bins) correspondientes a los ángulos 0, 20, 40,...,180 grados, como se muestra en la figura 3.6.

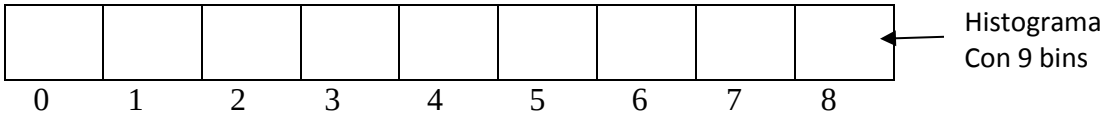


Figura 3.6. Histograma con 9 bins para almacenar información de gradientes con ángulos de 0 a 180 grados.

La figura 3.7 ilustra el proceso de cómo los valores de magnitud y orientación del gradiente intervienen en la formación del histograma. Se muestra la magnitud y dirección del gradiente de la sección de imagen de 8 × 8 pixeles obtenidas de la figura 3.4. Se selecciona un bin en función de la dirección, y el voto (el valor que entra en el bin) se elige de acuerdo al valor de la magnitud. En la figura 3.7 se observa que los píxeles con círculo azul expresan respectivamente los valores de 80 y 2, correspondientes a valores de dirección (ángulo en grados) y la magnitud del gradiente. Para almacenar estos valores en el histograma se ubica el intervalo que contenga al valor de 80, en este caso corresponde al bin 5, posteriormente se almacena en este bin el valor correspondiente a la magnitud del gradiente, 2 en este ejemplo. El gradiente en el píxel rodeado con rojo de la misma figura tiene un ángulo de 10 grados y una magnitud de 4. Dado que 10 grados se encuentra como valor frontera entre los bins 1 y 2, este valor se divide entre los dos intervalos, tal como se muestra en la figura 3.7, en este caso se almacena un valor de 2 en cada bin.

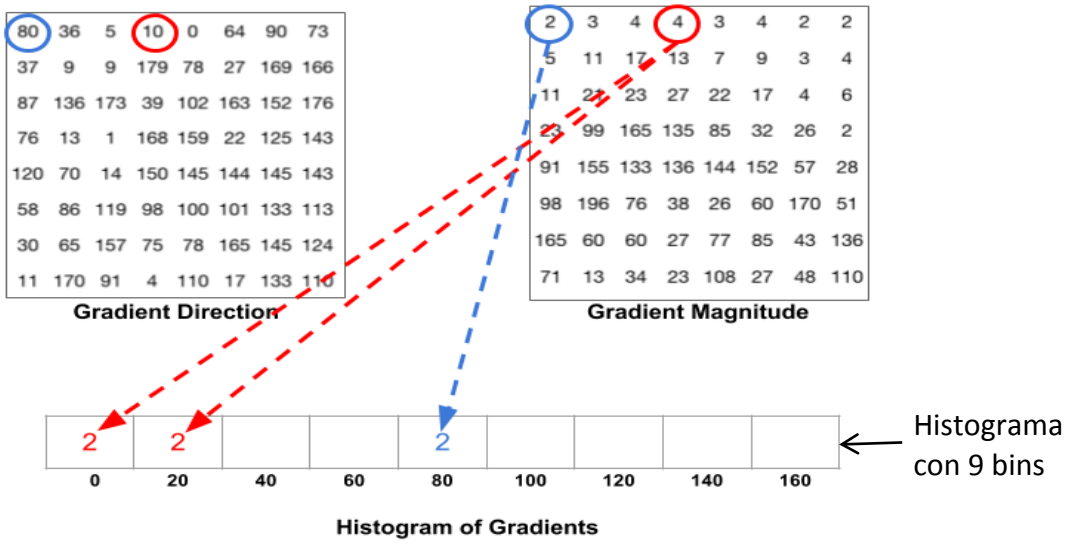


Figura 3.7. Histograma compuesto de 9 bins, que almacena la magnitud y dirección del gradiente de 0 a 160 grados de una sección de imagen de 8 x 8 pixeles [52].

Se debe tener en cuenta otro punto muy importante. Si el ángulo es mayor a 160 grados, consideremos que está entre 160^0 y 180^0, y sabemos que el ángulo se ajusta dando 0^0 y 180^0 equivalentes. Por lo tanto, un ángulo de 165^0 (figura 3.8, valor encerrado en círculo) se distribuye de forma proporcional entre los bins que almacena los ángulos contenidos de 0 grados y de 160 grados.

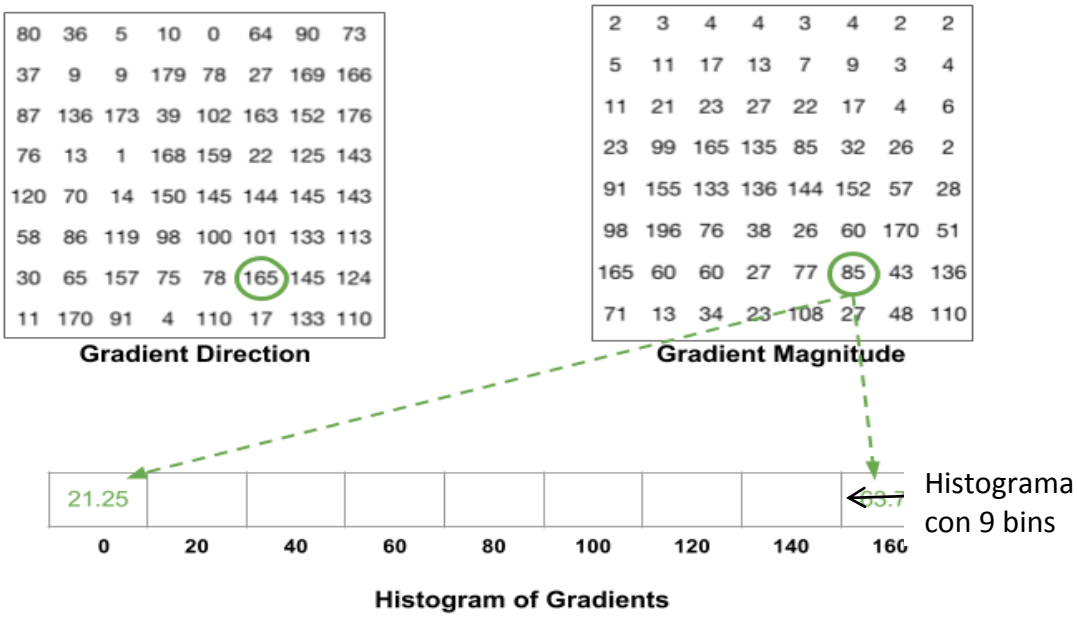


Figura 3.8. Histograma con magnitud y dirección del gradiente mayor a 160 grados [52].

Las contribuciones de todos los píxeles en las celdas de 8 × 8 pixeles se suman para crear el histograma de 9 bins, de cada una de las secciones de la imagen.

Los gradientes de una imagen son sensibles a la iluminación. Si se oscurece la imagen dividiendo todos los valores de píxel por 2, la magnitud del gradiente cambiará a la mitad, y por lo tanto los valores del histograma cambiarán a la mitad. Idealmente, queremos que el descriptor sea independiente de las variaciones de iluminación. En otras palabras, se tiene que "normalizar" el histograma para que sea robusto a los cambios de iluminación, para que ofrezca mayor resistencia al ruido y también para que se pueda manipular el tamaño de la imagen con mayor facilidad después de que ésta sea normalizada. El procedimiento de normalización se describe a continuación.

Suponga que se tiene píxel en formato de color RGB, con los siguientes valores [128, 64, 32]. La longitud de este vector se calcula como sigue $\sqrt{128^2 + 64^2 + 32^2}$ = 146.64. Esto también se llama la norma L2 del vector. Dividir cada elemento de este vector por el valor 146.64 nos proporciona un vector normalizado con los valores [0.87, 0.43, 0.22]. Ahora consideremos otro vector con los elementos que son dos veces el valor del primer vector como se muestra 2 x [128, 64, 32] = [256, 128, 64]. Esto se resuelve de la misma forma que el primer vector, dando como resultado el valor de 293.284. Dividiendo cada elemento del vector [256, 128, 64] por el valor 293.284 dará como resultado [0.87, 0.43, 0.22], que es lo mismo que la versión normalizada del vector RGB original. Se puede notar que la normalización de un vector lo hace invariante a los cambios de escala, en este caso a los cambios de iluminación.

Ahora que se sabe cómo normalizar un vector, se puede pensar que al calcular el vector HOG simplemente puede normalizar el histograma 9 × 1 de la misma manera que normalizamos el vector 3 × 1. No es una mala idea, pero una mejor idea es normalizar en un bloque de 16 × 16 pixeles de mayor tamaño como se muestra en la figura 3.9. Un bloque de 16 × 16 pixeles, tiene 4 histogramas (generados por las cuatro celdas de 8 x 8) que se pueden concatenar para formar un vector de 36 x 1 elementos, y se puede normalizar de la misma manera que se normaliza un vector de 3 × 1. Pero, entonces la ventana se mueve luego en 8 píxeles y se calcula un vector normalizado de 36 × 1 sobre esta ventana y el proceso se repite.

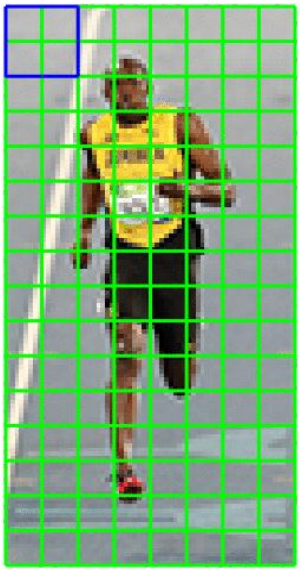


Figura 3.9. Bloque de 16 x 16 pixeles (cuatro celdas) [52].

Para calcular el vector de características final de la sección de la imagen completa, los vectores de 36 × 1 se concatenan en un vector gigante. ¿Cuál es el tamaño de este vector?

3.4 Cálculo del Vector

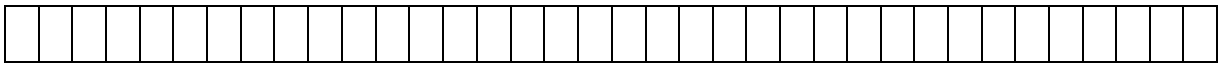
1. **¿Cuántas posiciones de los bloques de 16 × 16 se tienen en la imagen 3.9?**

En la imagen se observan 7 posiciones horizontales y 15 posiciones verticales que hacen un total de 7 x 15 = 105 posiciones.

2. Cada bloque de 16 × 16 está representado por un vector de 36 × 1. Entonces cuando los concatenamos a todos en un vector gigante obtenemos un vector 36 × 105 = 3780 datos

El descriptor HOG de una sección de imagen generalmente se visualiza trazando los histogramas normalizados 9 × 1 en las celdas de 8 × 8 pixeles y se pueden notar más visiblemente en los costados de la figura 3.11.

a).-Concatenación de un bloque de 16 x 16 en un gradiente de 4 x 1



b).-Concatenación de un bloque de 16 x 16 en un gradiente de 36 x 1

Figura 3.10. Las figuras a y b muestran los gradientes representados por bloque de celdas de 16 x 16 pixeles.

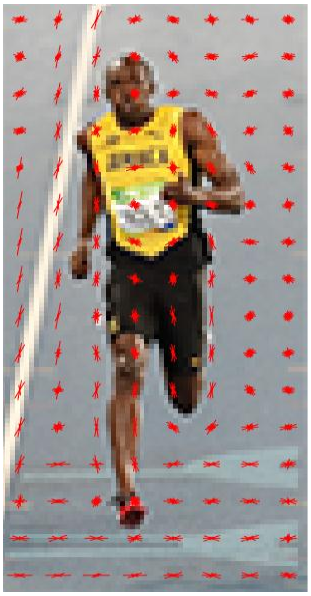


Figura 3.11. Visualización de histograma de gradientes orientados [52].

Para tener en cuenta los cambios en la iluminación y el contraste, las intensidades del gradiente deben estar normalizadas localmente, lo que requiere agrupar las celdas en bloques más grandes conectados espacialmente. El descriptor HOG es el vector concatenado de los componentes de los histogramas de celdas normalizados de todas las regiones de bloques. Estos bloques típicamente se superponen, lo que significa que cada celda contribuye

más de una vez al descriptor final. Existen dos geometrías de bloques principales: bloques rectangulares R-HOG y bloques circulares C-HOG. Los bloques R-HOG son generalmente cuadrículas, representadas por tres parámetros: el número de celdas por bloque, el número de píxeles por celda y el número de canales por histograma de celdas. En el experimento de detección humana [14], los parámetros óptimos fueron cuatro celdas de 8 x 8 píxeles por bloque (16 x 16 píxeles por bloque) con 9 canales de histograma. Además, descubrieron que se podía obtener una pequeña mejora en el rendimiento aplicando una ventana espacial gaussiana dentro de cada bloque antes de tabular los votos del histograma para ponderar los píxeles alrededor del borde de los bloques.

Los bloques R-HOG parecen bastante similares a los descriptores de transformación de característica invariante de escala (SIFT). Los bloques circulares de HOG (C-HOG) se pueden encontrar en dos variantes: aquellos con una sola celda central y aquellos con una celda central dividida angularmente. Además, estos bloques C-HOG se pueden describir con cuatro parámetros: el número de contenedores angulares y radiales, el radio del contenedor central y el factor de expansión para el radio de los contenedores radiales adicionales.

En los experimentos realizados [14], encontraron que las dos variantes principales proporcionaban el mismo rendimiento, y que dos cajones radiales con cuatro compartimentos angulares, un radio central de 4 píxeles y un factor de expansión de 2 proporcionaban el mejor rendimiento en su experimentación. Además, la ponderación Gaussiana no proporcionó ningún beneficio cuando se usa junto con los bloques C-HOG. Los bloques C-HOG parecen similares a los descriptores de contexto de forma , pero difieren mucho en que los bloques C-HOG contienen celdas con varios canales de orientación, mientras que los

contextos de forma solo utilizan un conteo de presencia de borde único en su formulación [14].

3.5 Máquinas de Vectores de Soporte

Las Máquinas de Vectores de Soporte (MVS) son modelos y algoritmos de aprendizaje supervisado que clasifican de manera confiable a un vector de características (nunca antes visto) en una clase determinada. La investigación que se está realizando para el reconocimiento de las expresiones faciales en este trabajo, aborda tres etapas que se plantean resolver para llegar a la solución.

La primera etapa comprende realizar la detección del rostro en las imágenes o videos utilizados para llevar a cabo la detección de las expresiones faciales, utilizando la imagen integral propuesta por de Viola y Jones [83] y la técnica de *Boosting* introducida por Schapire y sus colegas [35], para la utilización de los filtros de Haar. La segunda etapa comprende la extracción de características en las imágenes, para realizar este proceso se utilizará la técnica de Histogramas de Gradientes Orientados (HOG) [15], en combinación con las Máquinas de Vectores de Soporte (MVS), para realizar la clasificación de las características de las expresiones faciales. La tercera etapa comprende mostrar los resultados de las características detectadas en cada una de las imágenes, aplicando la Máquina de Vectores de Soporte a cada uno de los vectores detectados en las imágenes.

La aplicación de la técnica de los Histogramas de Gradientes Orientados (HOG), que se utiliza para detectar las características de las expresiones faciales en combinación con las Máquinas de Vectores de Soporte (MVS) [11], es una herramienta que se pretende utilizar

aplicando los métodos y técnicas para demostrar que resulta eficiente su desempeño para llevar a cabo su aplicación. La detección de las expresiones faciales en una persona es un problema que no está del todo resuelto debido a que no se sabe con exactitud qué es lo que las provoca cuando se presentan, pero que de alguna forma se manifiestan en el rostro del individuo mostrando algún estado emocional a través de éstas.

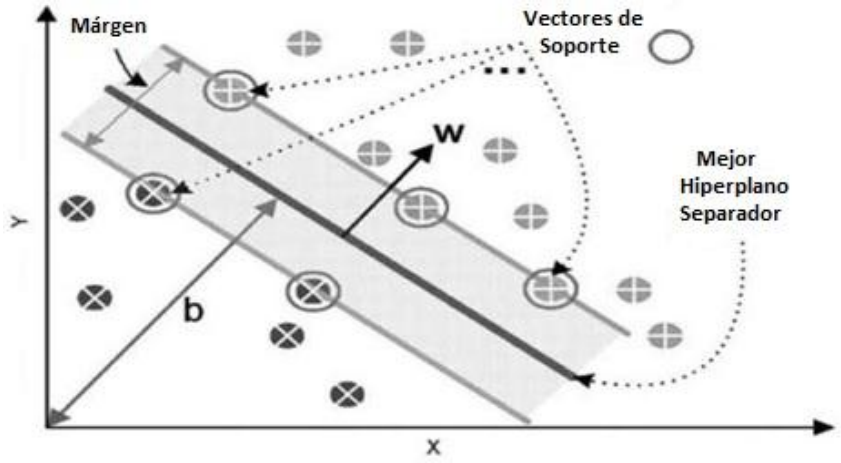


Figura 3.12 Separación con un Kernel Lineal [33].

3.5.1 Clasificación de Sitios Potenciales

Las máquinas de vectores de soporte (MVS), son un conjunto de técnicas o de algoritmos de aprendizaje supervisado. Estos métodos están propiamente relacionados para resolver problemas de clasificación y regresión.

Dado un conjunto de muestras de entrenamiento se pueden etiquetar las clases y entrenar un modelo de una MVS para construir un modelo que prediga la clase de una nueva muestra. Intuitivamente, una MVS es un modelo que representa a los puntos de muestra en el espacio, separando las clases en dos espacios lo más amplios posibles mediante un hiperplano de

separación definido como el vector entre los dos puntos, de las dos clases, más cercanos al que se llama vector de soporte. Cuando las nuevas muestras se ponen en correspondencia con dicho modelo, en función de los espacios a los que pertenezcan, pueden ser clasificadas en una u otra clase.

3.5.2 Clasificador Basado en Máquina de Vectores de Soporte (MVS)

La siguiente etapa consiste en emplear un clasificador que ayude a eliminar los falsos positivos producidos en la fase de detección. Para este fin, se utiliza la técnica conocida como Máquina de Vectores de Soporte (MVS) [11]. Las MVS han sido desarrolladas como una técnica robusta aplicadas en tareas de clasificación y regresión, y tienen la ventaja, de que pueden ser utilizadas para resolver tanto problemas lineales como no lineales. La idea principal consiste en construir un hiperplano de separación entre clases, de tal manera que el margen de separación entre las mismas sea máximo (hiperplano óptimo).

Para la construcción del hiperplano se debe considerar un conjunto de entrenamiento especificado por la ecuación $\{x_i,\ d_i\}_{i=1}^{N}$ donde $\boldsymbol{x_i}$ representa el valor de entrada y $\boldsymbol{d_i}$ su correspondiente valor objetivo, es decir, un proceso de aprendizaje supervisado. La distancia entre un hiperplano y el punto de los datos más cercano al mismo se denomina margen de separación $\boldsymbol{\rho}$. La frontera de decisión se especifica por la siguiente expresión.

$$w^T.x - b = 0 \qquad 3.3$$

Donde ***x*** representa el vector de entrada, ***w*** el vector de pesos ajustable y ***b*** el sesgo. Cuando se intenta obtener el hiperplano óptimo, la función discriminante viene dada por la Ec. siguiente.

$$g(x) = w_0^T . x - b_0 \qquad 3.4$$

A los puntos $\{x_i,\ d_i\}$ que se encuentran más cercanos al hiperplano óptimo y cuyo margen de separación es máximo, se les denominan vectores de soporte. Estos puntos resultan difíciles de clasificar ya que se encuentran en el límite de separación de las clases. En la construcción del hiperplano óptimo se emplean los multiplicadores de Lagrange. Por lo tanto, el problema de optimización se plantea como sigue: dados los datos de entrenamiento $\{x_i,\ d_i\}_{i=1}^{N}$, encontrar los valores óptimos del vector de pesos *w* y de sesgo *b* que satisfagan el siguiente criterio:

$$d_i(w^T . x_i) + b \ \geq 1 \ \text{ para } i = 1,2,3,\dots,N \qquad 3.5$$

Así mismo, se define el vector de pesos que minimice la siguiente función de costo:

$$\Phi(w) = \frac{1}{2} w^T . w \qquad 3.6$$

El problema de optimización a resolver se define por un modelo de programación con restricciones, es decir:

$$Q(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j K(x_i, x_j) \qquad 3.7$$

Sujeto a las restricciones:

$$\sum_{i=1}^{N} \alpha_i d_i = 0 \qquad 3.8$$

Con $\alpha_i \geq \alpha_j$ para $i = 1, 2, 3, \ldots, N$

Donde $K(x_i, x_j)$ es la función kernel[7] utilizada. En este caso, una función lineal.

La forma para resolver este problema de optimización es aplicando los multiplicadores de *Lagrange*, debido a que de esta manera es posible representar el hiperplano buscado como una combinación lineal de los mismos datos, y no en términos de la base del espacio vectorial en que aparecen. El hiperplano encontrado con la máquina de vectores de soporte representa el punto de separación entre las clases de información de los valores positivos y negativos que representan el rostro detectado en la imagen.

[7] Una función núcleo o kernel es un producto interno en el espacio de características que contiene su equivalente en el espacio de entrada.

Capítulo 4 Metodología

4.1 Introducción

En este trabajo se presenta una arquitectura basada en la técnica de HOG para extraer y representar las características de interés en una imagen y las MVS para clasificar de manera apropiada estas características, aun cuando no se tengan datos suficientes para el entrenamiento. La intención consiste en desarrollar una herramienta que reconozca de manera confiable una de las cuatro emociones humanas básicas expresadas por una persona en ambientes no controlados.

A continuación, se describe la metodología utilizada.

4.2 Metodología Utilizada

La metodología utilizada en este trabajo con el fin de reconocer de manera automática algunas de las emociones humanas básicas analizadas y estudiadas por Ekman [21] y Pantic [59] se muestra en la figura 4.1. A continuación se describe cada etapa.

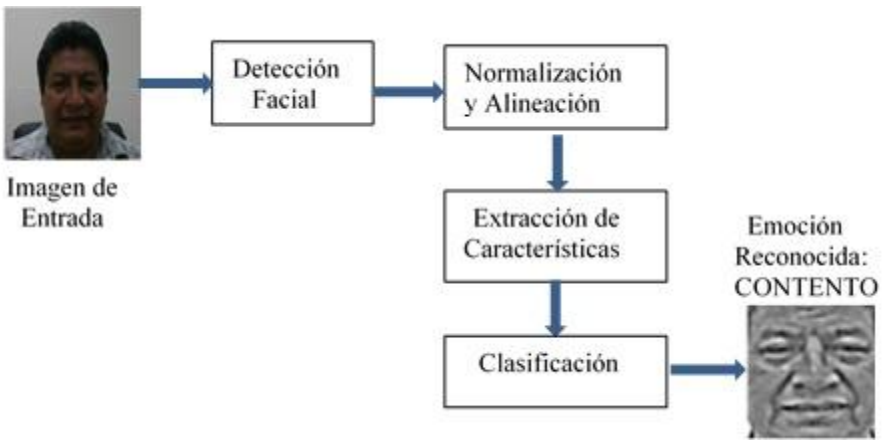


Figura 4.1. Etapas para el reconocimiento de expresiones faciales.

4.3 Imagen de Entrada

Las imágenes utilizadas en este trabajo, se capturan a través de la cámara web de la computadora y se almacenan con un tamaño de 640 x 480 pixeles (Figura 4.2). Estas imágenes son de manera inicial mejoradas con el fin de resaltar ciertas características (bordes, orillas) para un procesamiento posterior. Ya que posteriormente se pretende analizar el rendimiento del algoritmo considerando algunas técnicas de mejoramiento, se van a manejar tres aspectos: Imagen de entrada sin tratamiento, imagen de entrada ecualizada e imagen de entrada aplicándole un filtro de altas frecuencias (Sobel).

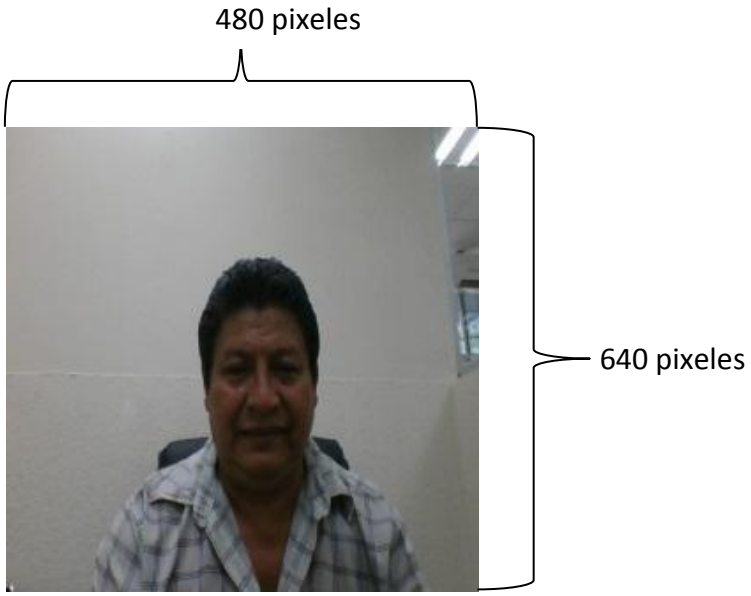


Figura 4.2. Imagen de entrada

4.3.1 Imagen Integral

La detección del rostro se lleva a cabo aplicando el concepto de imagen integral implementada en el algoritmo desarrollado por Viola y Jones [83]. Este algoritmo utiliza una imagen integral para extraer características de forma rápida y precisa, debido a que no trabaja

directamente con los valores de intensidad de los pixeles, sino que lo hace a través de una imagen acumulativa que se va formando basada en las operaciones básicas que se realizan a medida que se va recorriendo la imagen. La figura 4.3 ilustra la aplicación de este proceso con el fin de obtener la imagen integral (recuadro superior derecho) a partir de la imagen original Im(x, y). La imagen integral se obtiene al realizar un desplazamiento de izquierda a derecha y de arriba hacia abajo realizando la suma de los pixeles a medida que se va desplazando en la localización de los puntos (x, y), como se muestra en la figura 4.3 aplicando la Ec. 4.1.

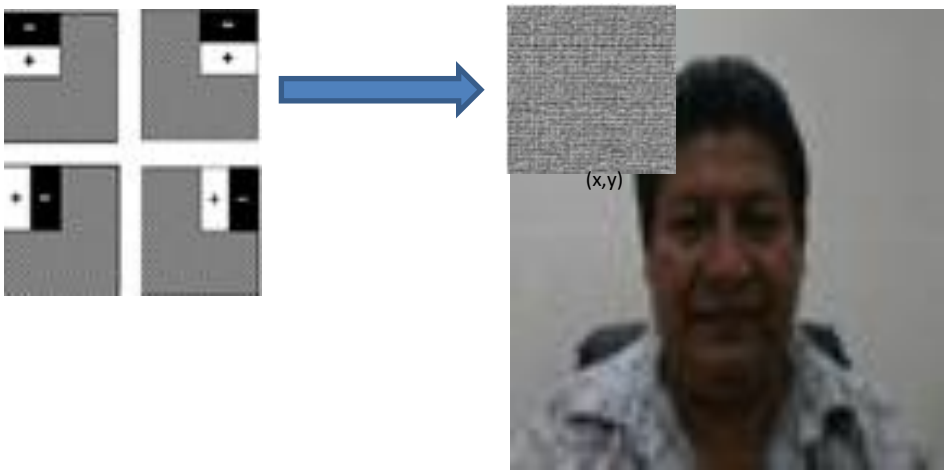


Figura 4.3. Operaciones básicas aplicadas a la imagen original.

En la figura 4.4 se muestran los filtros Haar que son utilizados para realizar la codificación de diferencia de intensidades en la imagen. La forma en que estas operaciones básicas se van aplicando se ilustran en la imagen, en la cual se muestran tres operaciones diferentes: suma y resta entre filas, suma y resta entre columnas, suma y resta en diagonal.

Estas operaciones se realizan a través de un proceso de filtrado (Filtros de Haar) o convolución sobre toda la imagen aplicando la Ec. 4.1

Figura 4.4. Convolución de filtros para la detección del rostro.

$$II(X,Y) = \sum_{x'\leq x;y'\leq y} Im(X',Y') \qquad 4.1$$

Donde:

II(X,Y).-Representa la imagen integral

Im (X',Y').-Representa la imagen original

Condiciones:

x'.-Menor o Igual a x

y'.-Menor o Igual a y

4.3.2 Extracción de Características para la Identificación Facial

La extracción de características es un paso importante en el reconocimiento de patrones debido a que cuando éstas son procesadas proporcionan atributos importantes que son utilizados para representar a los objetos de tal manera que se pueden agrupar en una determinada clase. En este caso, estas clases serán los atributos que describan el rostro y

atributos que no describan el rostro, de acuerdo a las clases de características (patrones) reconocidas.

En el reconocimiento facial la extracción de características es aplicada a la imagen utilizando los filtros con bases Haar. Estos filtros son calculados eficientemente sobre la imagen integral y son selectivos en la orientación espacial y la frecuencia, además permiten ser modificados en escala y orientación de acuerdo a las necesidades requeridas, es decir, si se requiere agrandar la imagen se utiliza un múltiplo y si se requiere minimizar se utiliza un divisor en la escala. Cuando se aplican los filtros Haar, éstos realizan una codificación de diferencia de intensidades en la imagen y no en los pixeles que contiene debido a que éstos trabajan con valores (0, 255), generando características de contornos, puntos y líneas, mediante la captura de contraste entre las regiones donde se aplican. Una vez que se obtienen las características de una imagen original, representadas a través de la imagen integral, el siguiente paso consiste en introducirlas a un clasificador con la finalidad de caracterizar la información, en nuestro caso el rostro de una persona.

4.3.3 Identificación del Rostro

La técnica de *Boosting* introducida por Schapire y Freund [35] consiste en un método de clasificación que utiliza varios clasificadores básicos para formar un único clasificador más complejo y preciso. Esta técnica se basa en el hecho de que varios clasificadores sencillos pueden combinarse para formar un clasificador que sea de mejor precisión, siempre y cuando se disponga de un número suficiente de muestras de entrenamiento. La aplicación de clasificadores en cascada ha permitido obtener buenos resultados en las muestras de entrenamiento, entre mayor sea el número de muestras, habrá mayor precisión en los

resultados obtenidos, como se demuestra en los trabajos realizados por Viola y Jones [83], [63].

Para aplicar la técnica de *Boosting* primero se debe establecer un algoritmo de aprendizaje sencillo (clasificador débil o base), que será llamado repetidas veces para crear diversos clasificadores base. Para el entrenamiento de los clasificadores base se emplea en cada iteración, un subconjunto diferente de muestras de entrenamiento y una distribución de pesos diferente sobre las muestras [35]. Entre mayor sea el número de muestras de entrenamiento mayor será la precisión en la clasificación de las características. Finalmente, estos clasificadores base se combinan en un único clasificador que es mucho más preciso que cualquiera de los clasificadores base por separado. En este trabajo se utilizaron 37 muestras de imágenes originales mostrando el mismo gesto/estado emocional de una misma persona, por lo tanto, se utilizaron un total de 128 imágenes para el entrenamiento del clasificador. En la Figura 4.5 se observa la identificación del rostro aplicando este algoritmo [43].

Figura 4.5 Detección del rostro aplicando el concepto de imagen integral y clasificadores en cascada (algoritmo de Viola & Jones).

4.4 Normalización y Alineación

Una vez que se tiene detectado el rostro de una persona en la imagen, el siguiente paso consiste en reconocer el estado emocional que muestra. Debido a que este proceso consiste en analizar algunas características presentes en la imagen y considerar puntos de referencia, es necesario normalizar y alinear la imagen con el fin de facilitar esta tarea. Se trata de ajustar la forma, tamaño y posición del rostro en la imagen con la finalidad de minimizar la variación en imágenes de la misma clase (sonriente, enojado, etc.). Para esto se aplican algunas funciones (filtros) estadísticas basadas en correlación. Las diferencias que existen entre estos filtros radican en cómo se construyen para procesar las muestras de entrenamiento que serán analizadas [43].

Los ejemplos incluyen Funciones Discriminantes Sintéticas (SDF) [39], filtros de Funciones Discriminantes Sintéticas de Varianza Mínima (MVSDF) [81], filtros de Energía de Correlación de Promedio Mínimo (MACE) [51], Filtros Óptimos de Compensación (OTF) [70] y Sin restricciones MACE (UMACE) [74].

Considerando que estos enfoques amplían en gran medida el rango de rendimiento de los filtros de correlación, todavía hay margen de mejora. Específicamente en este trabajo, se propone la utilización de la clase de filtros denominada Promedio de Filtros Sintéticos Exactos (ASEF) que difieren de los métodos anteriores en dos aspectos importantes [43].

En primer lugar, se especifica una superficie de respuesta de correlación completa para cada instancia de entrenamiento durante la construcción del filtro.

En segundo lugar, el resultado de los filtros utilizados en cada imagen de entrenamiento se promedia para obtener un mejor rendimiento [4].

La identificación y el registro preciso de las imágenes es un paso importante que se realiza en el reconocimiento facial, y un punto importante de establecer este registro facial es encontrar donde se encuentran ubicados exactamente los ojos en el rostro. Como se verá a continuación:

4.4.1 Localización de los Ojos

Generalmente los algoritmos de búsqueda de ojos utilizan las coordenadas (x, y) para ubicar los píxeles del centro de los ojos izquierdo y derecho en las imágenes frontales. Para que esto sea verdadero, el algoritmo debe devolver la ubicación del ojo proporcionando cierta tolerancia, medida típicamente como una fracción de la distancia interocular, es decir, la distancia entre los centros de los ojos en el rostro.

Se han desarrollado una gran cantidad de trabajos sobre la detección de los ojos utilizando las técnicas de los filtros. En este trabajo solo se mencionarán algunos porque también se usaron filtros de correlación sintéticos. Un algoritmo que se utiliza para la detección de rostros, es el clasificador en cascada de Viola y Jones [83], este trabajo ha sido adoptado por muchos investigadores para la detección de los ojos como se puede ver en [50] y [83]. Uno de los sistemas diseñados, usa un clasificador en cascada para el análisis del tono de la piel para la clasificación de las características [8].

En términos de filtros de correlación sintética y hallazgos oculares, el único trabajo previo que se tiene conocimiento es el de Brunelli y Poggio en 1997. En este trabajo se aplicó

correlación de filtros para detección de ojos Filtros Sintéticos Discriminantes (SDF) y Funciones Discriminantes Sintéticas de Mínimos Cuadrados (LSSDF) en la detección ocular [7].

Después de realizar la captura de la imagen y aplicar la técnica de la imagen integral para detectar el rostro, se aplica el método del Promedio de Filtros Sintéticos Efectivos (ASEF) [4], para ubicar los puntos referenciales para la mejor ubicación del rostro, la imagen queda en escala de grises como se muestra en la figura 4.6.

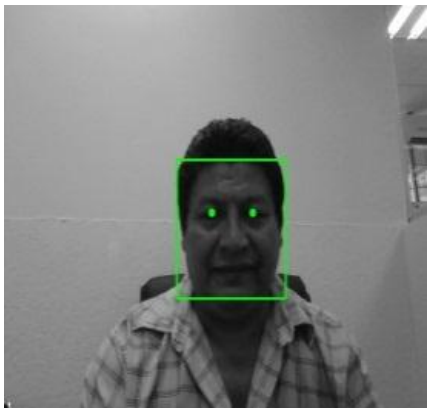


Figura 4.6. Imagen después de aplicar la técnica de ASEF para la detección de los ojos.

4.4.2 Delimitación de la Imagen

Después de haber detectado el rostro en la imagen de entrada, el siguiente paso es delimitar el contorno facial de la imagen (rostro) para observar que están presentes todos y cada uno de los componentes faciales (ojos, boca, nariz, cejas, frente). Esto se realiza con la finalidad de dejar solamente el rostro de la imagen que se muestra en la figura 4.6, dejando a un lado los otros elementos componentes del rostro como: orejas, pelo, cuero cabelludo y cualquier

otro objeto (aretes o algún tatuaje), que pueda estar presente en el rostro y que no sean relevantes para la aplicación.

Este punto es muy importante debido a que la imagen debe estar completamente despejada de cualquier objeto que pueda proporcionar información inadecuada que interfiera en el siguiente proceso. El rostro delimitado y alineado durante la fase de preprocesamiento se muestra en la figura 4.7.

Figura 4.7. Imagen del rostro delimitado obtenido de la imagen de entrada. (Imagen original obtenida con la cámara web).

Hasta este punto se ha desarrollado la primera parte de esta investigación, como se muestra en la figura 4.7, ya se tiene un rostro delimitado.

Una vez que se tiene detectada la región de interés el siguiente paso consiste en extraer las características de esta región con el fin de identificar el vector de patrones que mejor representan a cada una de las emociones básicas que se pretenden reconocer. Para esta tarea se van a utilizar los histogramas de gradientes orientados.

4.5 Extracción de Características para la Clasificación de Emociones Básicas

El proceso de extraer y representar las características en la imagen delimitada se va a realizar a través de los Histogramas de Gradientes Orientados (HOG) los cuales se aplican para representar la apariencia local y la forma de un objeto en una imagen, esto se puede caracterizar como la distribución de intensidad del gradiente a través de la imagen, esto es, por las intensidades de los pixeles. Esta caracterización se obtiene dividiendo la imagen en regiones pequeñas y conectadas entre sí para formar grupos de celdas o bloques. A estos bloques se les calcula la variación de intensidad entre píxeles vecinos (celdas) los cuales van a formar los vectores de características o descriptores, estos valores van a ser normalizados.

Con la normalización del descriptor no se requiere normalizar cada una de las celdas dentro de los bloques, por lo tanto, la normalización llevada a cabo a través de bloques de celdas presenta una gran ventaja debido a que el resultado es el mismo que cuando se realiza en una celda, además la normalización hace que el descriptor sea menos sensible a los cambios de iluminación y al ruido.

4.5.1 Descriptor de Características

Un descriptor de características (figura 4.8) es una representación vectorial de una imagen o una sección de una imagen. Típicamente, un descriptor de características convierte una imagen 2D (r, g, b) a un vector (conjunto de características de longitud **n)**, en el descriptor de características HOG.

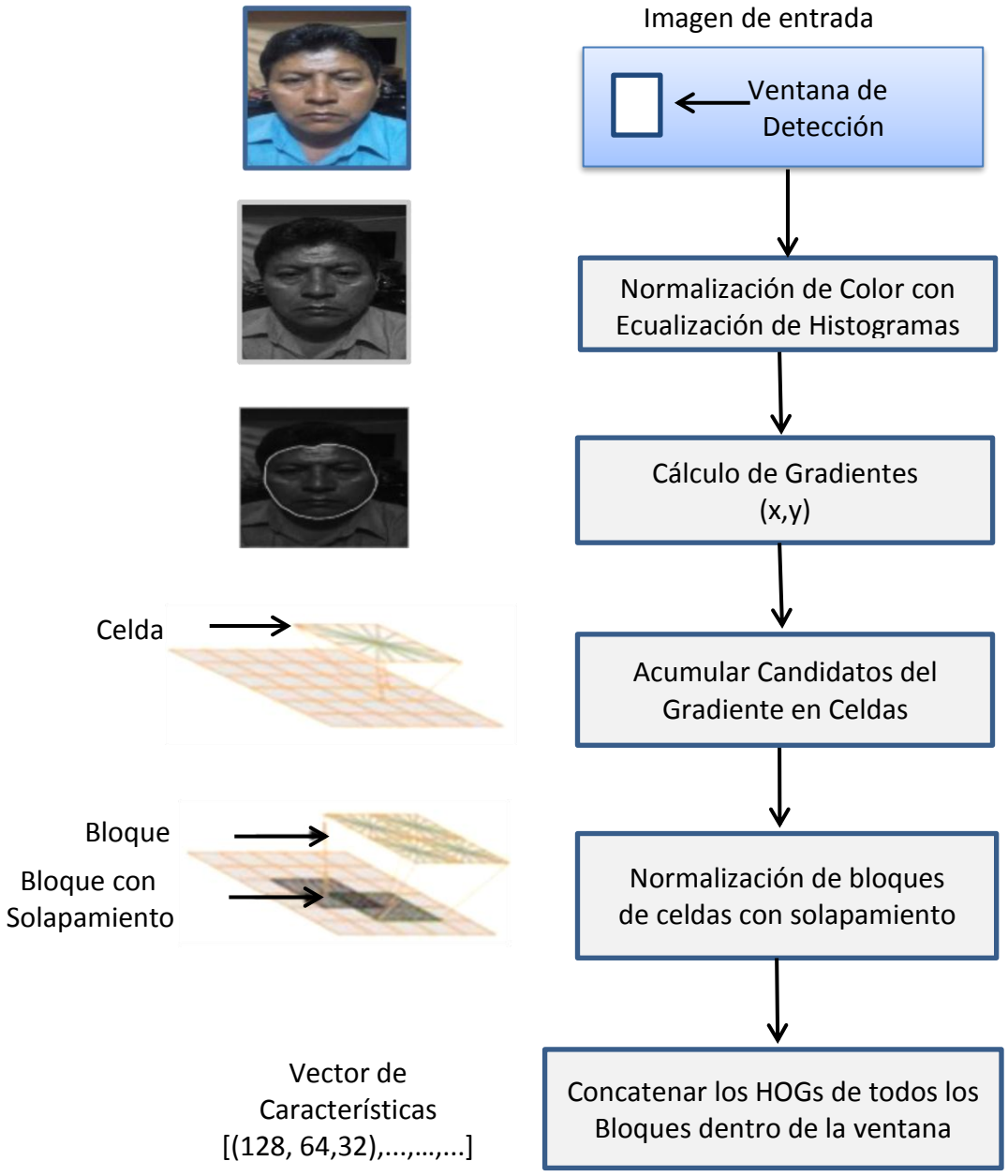


Figura 4.8 Procedimiento de bloques para la extracción del descriptor HOG en una ventana de detección de una imagen [10].

En este trabajo, la imagen de entrada es de tamaño 640 x 480 pixeles y después del preprocesamiento y al delimitarla es convertida a un tamaño 96 x 80 pixeles.

4.5.2 Descripción de la Imagen

Ventana de Detección. En el primer bloque se obtiene la imagen de entrada en la cual se detecta el rostro facial de una persona para ser analizada a través de los histogramas de gradientes orientados. El rostro debe estar libre de objetos que obstruyan la detección facial (sin aretes y sin lentes o cualquier otro objeto).

Normalización de Color con Ecualización de Histogramas. El segundo bloque se encarga de normalizar los colores de los pixeles contenidos en la imagen original (poner los bits en un solo color), esta normalización se logra aplicando la técnica de ecualización de histogramas para que la imagen pueda ser manipulada de mejor manera. Además, con la normalización realizada a las celdas, los pixeles se hacen menos sensibles al ruido. La ecualización tiene por objetivo mejorar la apariencia visual de la imagen, para esto, amplía el margen de los colores si es que se encuentran concentrados en un pequeño intervalo.

Cálculo de Gradientes. Después de haber normalizado la imagen de entrada se debe extraer información del rostro, para ello se utiliza el cálculo del gradiente, esto se realiza con el descriptor de HOG. Primero debemos calcular los gradientes horizontales y verticales, esto se realiza filtrando la imagen con los núcleos (-1, 0, 1) y (-1, 0, 1) y posteriormente encontrar la magnitud y dirección de dicho gradiente.

Acumular Candidatos del Gradiente en Celdas. Cada gradiente calculado en el proceso anterior tiene una magnitud y una dirección, por esta razón, para tener una mayor facilidad se calcula en celdas de 8 x 8 pixeles. Estas celdas se van acumulando cada una ellas en un

vector de longitud **n**. La magnitud se calcula con $G = \sqrt{g_x^2 + g_y^2}$ y la orientación se calcula con $\Theta = arc\tan(\frac{gy}{gx})$.

Normalización de Bloques de Celdas con Solapamiento. Debido a que normalizar una celda de una sección de imagen proporcionaba un vector de 9 x 1, obteniendo un vector **n**, se obtenía el mismo resultado que normalizar un grupo de cuatro celdas, es decir un vector de 9 x 4, se optó por normalizar grupos de cuatro celdas.

Concatenar los HOG's de Todos los Bloques Dentro de la Ventana. Se deben concatenar todos los histogramas de los bloques, para formar un vector de 36 x 1, para obtener un vector mayor con la suma de los grupos de celdas contenidos en la imagen. Si la imagen contiene 105 bloques o grupos de celdas, y cada bloque tiene un vector de 36 x 1, el vector final tendrá un valor de 36 x 105 = 3780 datos.

4.5.3 Normalización y Truncamiento

Sea **F** un mapa de características de nivel de píxeles para una imagen **w x h**. Sea **k > 0** un parámetro que especifique la longitud lateral de una región de imagen. Definimos una cuadrícula densa de "celdas" rectangulares y características agregadas de nivel de píxel para obtener un mapa de entidades basado en celdas **C**, con los vectores de características **C (i, j)** para la siguiente ecuación.

$$\mathbf{0 \leq i \leq [(w - 1) / k] \ \ y \ \ 0 \leq j \leq [(h - 1)/ k].} \qquad 4.2$$

Esta agregación proporciona cierta invariancia a pequeñas deformaciones y reduce el tamaño de un mapa de características.

El enfoque más simple para agregar características es asignar cada píxel **(x, y)** en una celda **([x / k], [y / k])** y definir el vector de características en una celda como la suma (o promedio) de las características de nivel de píxel en esa celda. En lugar de asignar cada píxel a una celda única, utilizamos un enfoque de "agrupamiento suave" en el que cada píxel contribuye a los vectores de características en el bloque de las celdas [15].

Los gradientes son invariables a los cambios de iluminación en su inclinación. Esto se debe a que la invarianza en la ganancia se puede lograr a través de la normalización del vector. La normalización de un vector en una celda de 8 x 8 pixeles tiene 9 dimensiones que corresponden a cada uno de los ángulos de inclinación que puede tener la celda en el rango de 0^0 a 180^0.

En la investigación realizada por Dalal y Triggs [15], encontraron que se pueden definir vectores de características de baja dimensión que son sensibles al contraste. En dichos estudios se ha encontrado que los rendimientos en algunas categorías de objetos mejoran si usan características sensibles al contraste, mientras que algunas categorías se benefician del contraste de características insensibles al contraste. Por lo tanto, en la práctica, se utilizan vectores de características que incluyen ambos sensitivos y no sensitivos al contraste.

En este trabajo de investigación se utilizan 108 vectores dimensionales, definidos por 27 sumas sobre diferentes normalizaciones, uno para cada canal de orientación, **(9 insensibles al contraste y 18 sensibles al contraste) y 4 dimensiones que capturan la energía del gradiente en bloques de diez celdas (i, j). Por lo tanto, el mapa de características final tiene un vector de 31 dimensiones.**

El vector final de características se integra con los siguientes datos:

W = ancho de la imagen entre el tamaño de la celda = 80 / 8 = 10

H = alto de la imagen entre el tamaño de la celda = 96 / 8 = 12

HOG = dimensión del vector de características de HOG = 31

Por lo tanto, la fórmula utilizada es la siguiente:

Tamaño final del vector de características = W x H x HOG

Tamaño final del vector de características = 10 x 12 x 31 = 3720.

Después de haber aplicado la técnica de Histogramas de Gradientes Orientados (HOG) para el preprocesamiento, la imagen se recorta a un tamaño de 80 x 96 pixeles, y después se divide en celdas de 8 x 8 pixeles, y se forman bloques de 10 celdas que son utilizados para extraer vectores de características que se van acumulando en un vector final. Este vector almacenará la información que describirá de manera precisa todos los atributos que representarán la imagen analizada durante el preprocesamiento para encontrar la expresión facial en el rostro detectado, como se describe en la figura 4.9.

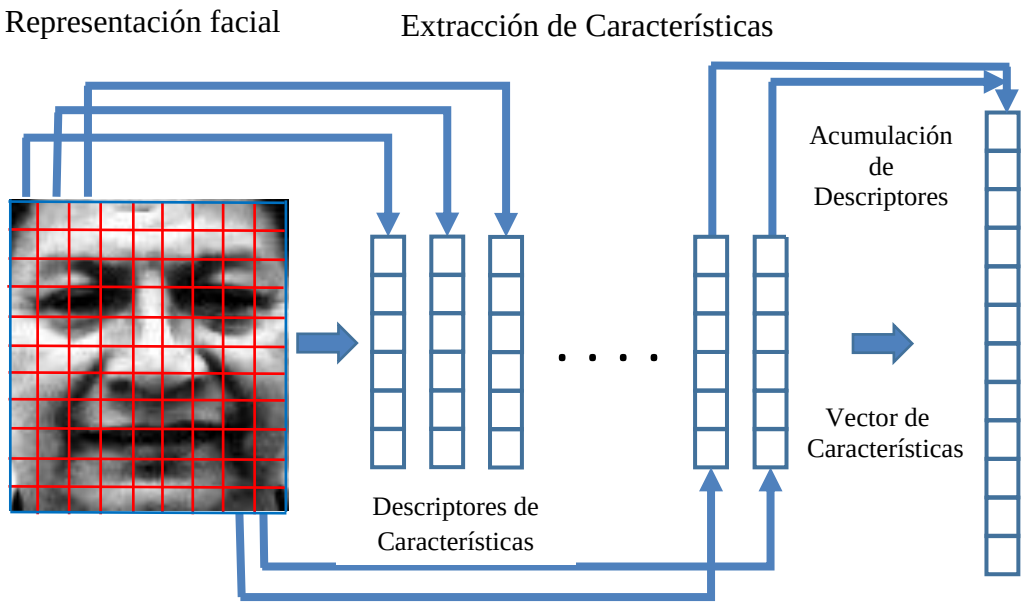


Figura 4.9. Rostro dividido en celdas para obtener los descriptores de características que se forman utilizando HOG, acumulados para formar el vector de características final.

La cantidad de vectores finales que se forman utilizando la técnica de HOG, para extraer las características dependerá del número de imágenes que se procesen. Los vectores obtenidos se introducirán en la máquina de vectores de soporte para hacer la clasificación en información útil e información no útil para el reconocimiento de las expresiones faciales.

4.6 Máquina de Vectores de Soporte

Una Máquina de Soporte Vectorial (MVS), es un sistema de aprendizaje automático que se ha utilizado para resolver problemas de clasificación y regresión de manera muy eficiente, debido a esto, ha sido reconocida como una de las mejores técnicas de clasificación, aún por

encima de otras técnicas de clasificación como las redes neuronales o los algoritmos genéticos.

Las MVS se basan en la Teoría de Aprendizaje Estadístico, su éxito se basa fundamentalmente en tres ventajas: La primera radica en que poseen una base matemática muy sólida; la segunda, en que se basan en el concepto de minimización del riesgo estructural, es decir, en minimizar el riesgo de una clasificación errónea al introducir nuevos datos como ejemplos y la tercera, en que disponen de potentes herramientas y algoritmos para hallar la solución de manera rápida y eficiente en un problema con una gran cantidad de datos de información [11].

De esta manera, estas máquinas son capaces de clasificar muestras en dos posibles conjuntos de información "positiva" y "negativa". Para realizar este proceso, se requiere de un entrenamiento previo de la máquina con esta información, por lo que se le introduce ejemplos de información "positiva" y "negativa", que corresponden a las imágenes faciales para que realice dicha clasificación.

En este caso, se trata de resolver un problema de clasificación linealmente separable, es decir, la máquina debe aprender una superficie de decisión adecuada, basada en los datos de entrenamiento. La superficie de decisión es un hiperplano que divide el conjunto de datos de entrenamiento en las dos clases. En ese sentido, el conjunto de datos de entrenamiento es de la forma (x_1, y_1), $(x_2, y_2),\ldots, (x_N, y_N)$, donde $\boldsymbol{xi} \in \mathbb{R}^d$ y corresponde al vector que se va a clasificar, en tanto que $yi \in \{0,1\}$ es la etiqueta correspondiente a cada vector.

4.6.1 Funcionamiento de la MVS

A continuación, se va a explicar funcionamiento de la MVS, en el que solo se tienen dos clases; generalmente llamado clasificación binaria. Dado un conjunto **L** de puntos etiquetados **{x_i, y_i}** con $\boldsymbol{xi} \in \mathbb{R}^d$ es un vector de características o características del i-ésimo objeto y **yi ϵ {-1, +1}** de la etiqueta de la clase, queremos construir una regla para determinar, dada una nueva **x**, una de las dos clases posibles. También asumiremos que nuestro dato es linealmente separable, es decir que podemos dibujar una línea que divide todos los puntos pertenecientes a una clase de los puntos que pertenecen a la otra clase cuando **d = 2**, y un hiperplano cuando **d > 2**. Además, queremos que este hiperplano de separación maximice los márgenes o las distancias a los puntos más cercanos de cada clase.

El hiperplano mencionado anteriormente puede describirse por **w. x + b = 0** donde **w** es normal al hiperplano y **b/‖w‖** es la distancia perpendicular desde el hiperplano al origen. Luego, resolver la clasificación puede escribirse como:

$$\mathbf{xi} \bullet \mathbf{w} + \boldsymbol{b} \leq \mathbf{1} \quad \textbf{para } \mathbf{y_i} = \mathbf{1} \tag{4.3}$$

$$\mathbf{xi} \bullet \mathbf{w} + \boldsymbol{b} \leq \mathbf{-1} \quad \textbf{para } \mathbf{y_i} = \mathbf{-1} \tag{4.4}$$

Estas dos ecuaciones se pueden describir con una sola ecuación en la siguiente forma:

$$\mathbf{yi}\ (\mathbf{x_i} \bullet \mathbf{w} + \boldsymbol{b}\ ) \mathbf{-1} \geq \mathbf{0}\ \forall \mathbf{i} \tag{4.5}$$

Si los conjuntos de puntos son linealmente separables, entonces se puede dibujar una línea o hiperplano pasando por los puntos que se encuentran más cerca del hiperplano de separación, esto es, los vectores de soporte.

El objetivo principal de Máquina de Vectores de Soporte (MVS) es maximizar la distancia entre los nuevos hiperplanos; referenciados como H_1 y H_2. La Figura 4.1 ilustra la teoría explicada anteriormente.

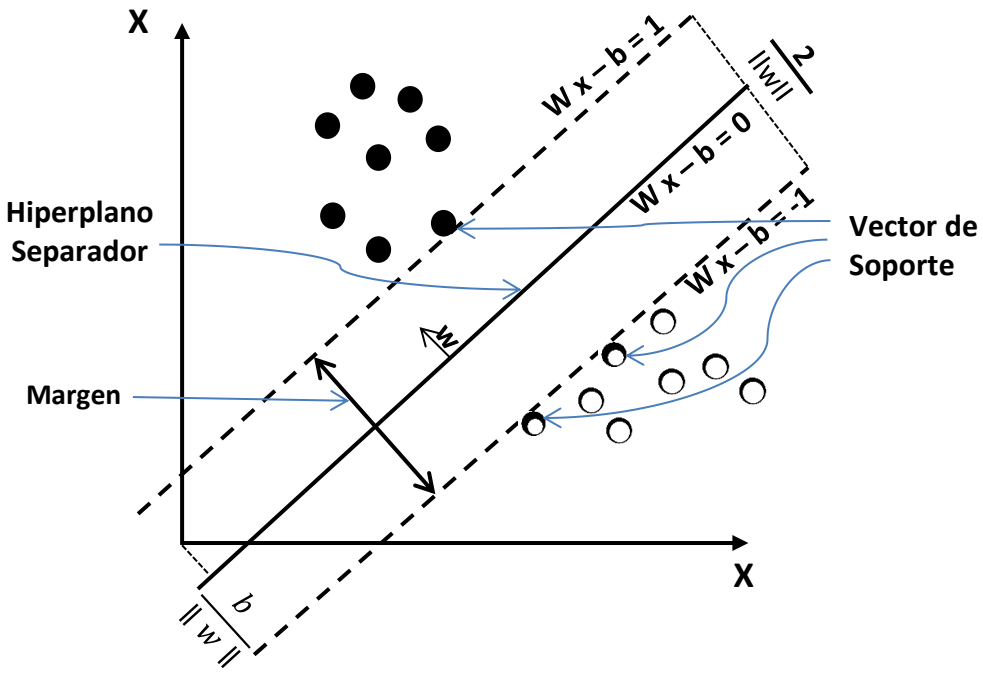


Figura 4.10. Margen del hiperplano de separación máxima de la Máquina de Vectores de Soporte [72].

Geométricamente se puede ver que la distancia de H_1 al hiperplano es igual a $\frac{1}{\|w\|}$ y de manera similar a H_2, por lo que la distancia de H_1 a H_2 es igual a $\frac{2}{\|w\|}$. Como explicamos que el objetivo principal es maximizar esta distancia, entonces tenemos que resolver el problema se reduce para minimizar $\|w\|$. Minimizar $\|w\|$ es equivalente a minimizar $\frac{1}{2} \bullet \|w\|^2$ pero usando

el término posterior hace posible realizar la Optimización de Programación Cuadrática (QP). Por lo tanto, se tiene la ecuación:

$$\min_{(w,b)} \frac{1}{2} \bullet \|w\|^2$$

$$\text{s. t.} \quad y_i\,(x_i \bullet w + b) - 1 \geq 0 \;\forall i \qquad 4.6$$

Independientemente de los detalles de la programación cuadrática (QP) que se realice, la intuición detrás de la máquina de vectores de soporte (MVS) se puede explicar fácilmente con la figura 4.11, donde se muestran tres planos diferentes.

a).- El hiperplano $\mathbf{H_1}$ separa los puntos, pero no satisface la restricción de margen máximo.

b).- El hiperplano $\mathbf{H_2}$ separa ambos conjuntos y maximiza la distancia desde el hiperplano a los vectores de soporte.

c).- El hiperplano $\mathbf{H_3}$ no separa los componentes por completo de los dos conjuntos de puntos.

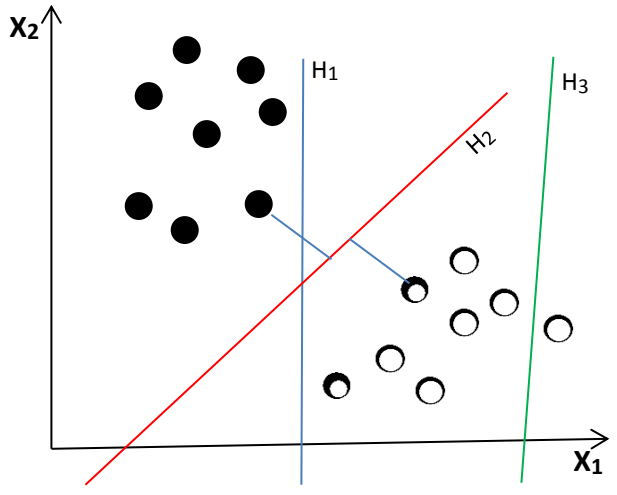


Figura 4.11. Diferentes hiperplanos en 2D: esta figura muestra tres separaciones diferentes, pero solo H2 proporciona el margen de separación máximo [72].

A cada uno de los vectores finales que se formaron con la suma de todos los descriptores de características de los bloques de celdas que componen la imagen se les asigna un número de etiqueta (correspondiente a cada imagen) para llevar a cabo su identificación y clasificación para que posteriormente alimenten el algoritmo de aprendizaje. Estas etiquetas, que se utilizan como parámetros, se introducen en el algoritmo de clasificación (Máquina de Vectores de Soporte-MVS) que se encarga de reconocer y clasificar los estados emocionales básicos detectados en la imagen a través de los vectores de características, mostrando que se ha reconocido un estado emocional registrado en el descriptor.

El proceso de alimentación de la Máquina de Vectores de Soporte con los parámetros (valores numéricos) que representan a cada uno de los vectores finales de características que describen los estados emocionales de las imágenes almacenadas en la base de datos, se realiza para clasificar a éstas con la finalidad de reconocer el estado emocional detectado en ella. En seguida se realiza el mismo procedimiento para detectar el estado emocional en la siguiente imagen, así sucesivamente hasta terminar con todas las imágenes que contiene la base de datos, de esta forma se realiza el entrenamiento con la máquina de vectores de soporte.

4.6.2 Características Utilizadas en el Entrenamiento de las MVS

A continuación, se describe el histograma de características de gradientes orientados (HOG) de 36-dimensiones tomando como base los datos extraídos de un bloque de celdas de 16 x 16 pixeles de Dalal y Triggs [14] y se introduce un conjunto de características de 13-dimensiones alternativas que capturan esencialmente la misma información (existen pequeñas diferencias de información entre los conjuntos de características de 36 y de 13 dimensiones, que resultan insignificantes para ser consideradas). Se ha encontrado que

aumentando el conjunto de características de baja dimensión para incluir características sensibles al contraste y características insensibles al contraste, lo que lleva a un vector de características de 31 dimensiones, mejora el rendimiento para la mayoría de las clases de los conjuntos de datos.

Los gradientes extraídos y posteriormente normalizados son invariables a los cambios de iluminación. Esto se debe a que la invarianza en la ganancia se puede lograr a través de la normalización. Los investigadores Dalal y Triggs [15] utilizaron cuatro factores de normalización diferentes para extraer el vector de características **C (i, j)**. Podemos escribir estos factores como $\mathbf{N_{\delta,\Upsilon}\ (i, j)}$ con $\boldsymbol{\delta,\Upsilon \in \{-1, 1\}}$. Estos factores de normalización se calculan con la Ec. 4.7.

$$\mathbf{N_{\delta,\Upsilon}\ (i, j) = (\|C(i, j)\|^2 + \|C(i + \delta, j\|^2 + \|C(i, j + \Upsilon)\|^2 + \|C(i + \delta, j + \Upsilon)\|^2)^{½}} \qquad 4.7$$

Cada factor mide la "energía de gradiente" en un bloque cuadrado de cuatro celdas que contiene los valores del vector **(i, j)**. Sea $\mathbf{T\alpha(V)}$ para denotar el truncado por componentes de un vector **V** por **α** (la i-ésima entrada en $\mathbf{T\alpha(V)}$ es el mínimo de la i-ésima entrada de **V** y **α**). El mapa de características de HOG (figura 4.12), es obtenido concatenando el resultado de normalizar el mapa de características basado en celdas **C** con respecto a cada factor de normalización seguido de truncamiento utilizando la Ec. 4.8.

$$C(i, j) = \begin{bmatrix} T_{\alpha}(C(i, j)/N_{-1,-1}(i, j)) \\ T_{\alpha}(C(i, j)/N_{+1,-1}(i, j)) \\ T_{\alpha}(C(i, j)/N_{+1,+1}(i, j)) \\ T_{\alpha}(C(i, j)/N_{-1,+1}(i, j)) \end{bmatrix} \quad 4.8$$

Las características del histograma de gradientes orientados (HOG) comúnmente utilizadas se definen usando **P = 9** orientaciones de gradiente insensibles al contraste (correspondientes a los bins de los ángulos entre 0^0 y 180^0), un tamaño de celda de **k = 8** y truncamiento **α = 0.2**. Esto conduce a un vector de características de 36 dimensiones. Usamos estos parámetros en el análisis que se describe a continuación.

En la figura 4.12 cada vector propio se muestra como una matriz de 4 por 9, de modo que cada fila corresponde a un factor de normalización y cada columna a un bin de orientación. Los valores propios se muestran encima de los vectores propios. El subespacio lineal abarcado por los 11 autovectores principales captura esencialmente toda la información en un vector de características. Observe cómo todos los vectores propios superiores son constantes a lo largo de cada columna o fila de la representación de la matriz [29].

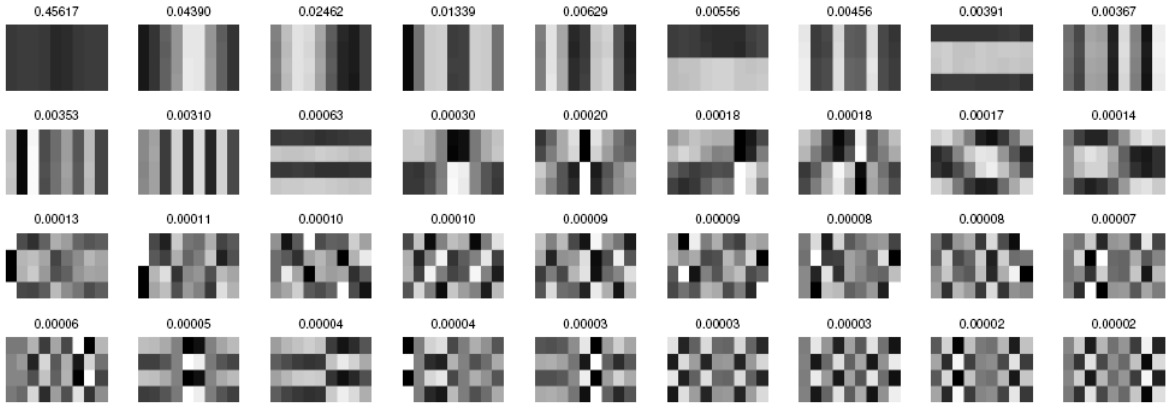


Figura 4.12 Funciones de PCA de HOG [29].

Un Histogramas de Gradientes Orientados de 36 dimensiones se define utilizando 4 normalizaciones diferentes de un histograma de 9 dimensiones sobre las orientaciones. Por lo tanto, una característica HOG de 36 dimensiones se ve como una matriz de 4 x 9.

Los vectores propios superiores en la Figura 4.12 tienen una estructura muy especial: cada uno (aproximadamente) es constante a lo largo de cada fila o columna de su representación matricial. Por lo tanto, los vectores propios superiores se encuentran (aproximadamente) en un subespacio lineal definido por vectores dispersos que tienen a lo largo de una sola fila o columna de su representación matricial.

A continuación, se muestra un ejemplo para calcular el vector dimensional de una matriz de 4 x 9, considerando que la proyección en cada $\mathbf{u_k}$ se puede obtener considerando las 4 normalizaciones para cada orientación fija, y la proyección en cada $\mathbf{v_k}$ también se puede obtener considerando más de 9 orientaciones para una normalización fija.

Sea $\mathbf{V = \{u_1,\ldots,u_9\}\ U\ \{v_1,\ldots,v_4\}}$ con

$$U_k(i, j) = \begin{cases} 1 & \text{si } j = k \\ 0 & \text{de otra forma} \end{cases} \qquad 4.9$$

$$V_k(i, j) = \begin{cases} 1 & \text{si } i = k \\ 0 & \text{de otra forma} \end{cases} \qquad 4.10$$

Podemos definir una característica de 13 dimensiones tomando el producto punto de una función de Histogramas de Gradientes Orientados de 36 dimensiones con cada $\mathbf{u_k}$ y $\mathbf{v_k}$. La proyección en cada $\mathbf{u_k}$ se calcula sumando las 4 normalizaciones para cada orientación fija, para este caso se utiliza la Ec. 4.9. La proyección en cada $\mathbf{v_k}$ se calcula sumando más de 9 orientaciones para una normalización fija, para este caso se utiliza la Ec. 4.10.

Sin embargo, el cálculo de las características de 13 dimensiones es mucho menos costoso que realizar proyecciones a los eigenvectores superiores obtenidos a través del método (PCA), ya que $\mathbf{u_k}$ y $\mathbf{v_k}$ son escasos. Además, las características de 13 dimensiones tienen una interpretación simple como 9 características de orientación y 4 características que reflejan la energía de gradiente general en diferentes áreas alrededor de una celda.

Sea **C** un mapa de funciones basado en celdas calculado al agregar un mapa de características con nivel de píxeles con **9** orientaciones insensibles al contraste. Sea **D** un mapa de funciones basado en celdas similar calculado utilizando **18** orientaciones sensibles al contraste. Definimos 4 factores de normalización para la celda **(i, j)** de **C** y **D** usando C como en la

ecuación (4.7). Podemos normalizar y truncar **C (i, j)** y **D (i, j)** usando estos factores para obtener **4 * (9 + 18) = 108** vectores de características dimensionales, **F(i, j)**. En la práctica, utilizamos una proyección analítica de estos 108-vectores dimensionales, definidos por 27 sumas sobre diferentes normalizaciones, uno para cada canal de orientación de **F**, y **4** sumas sobre las **9** orientaciones insensibles al contraste, uno para cada factor de normalización. Usamos un tamaño de celda de **k = 8** y el valor de truncamiento de **α = 0.2**. **El mapa de características final tiene vectores de 31 dimensiones G (i, j), con 27 dimensiones correspondientes a diferentes canales de orientación (9 insensibles al contraste y 18 sensibles al contraste) y 4 dimensiones que capturan la energía del gradiente general en bloques cuadrados de diez celdas (i, j).**

4.6.3 Entrenamiento de la MVS

Para la evaluación de este proyecto participaron 476 personas mostrando cada uno de ellos cuatro emociones básicas diferentes para la obtención de imágenes que fueron almacenadas en la base de datos.

Durante el entrenamiento de la Máquina de Vectores de Soporte, se utilizó el 50% de las imágenes obtenidas. Por lo tanto, se utilizaron 952 imágenes de rostros diferentes mostrando cada uno de ellos cuatro emociones básicas para el ensayo de la MVS. Se obtuvieron 1904 vectores de características extraídas del total de las imágenes almacenadas en la base de datos, las cuales como se mencionó anteriormente fueron escaladas de 640 x 480 pixeles a 80 x 96 pixeles.

El vector de características utilizado para el entrenamiento de la MVS, se compone de los siguientes datos:

W = ancho de la imagen entre el tamaño de la celda = 80 / 8 = 10

H = alto de la imagen entre el tamaño de la celda = 96 / 8 = 12

HOG = dimensión del vector de características de HOG = 31

Por lo tanto, la fórmula utilizada es la siguiente:

Tamaño final del vector de características = W x H x HOG

Tamaño final del vector de características = 10 x 12 x 31 = 3720.

Considerando que se obtuvieron 1904 imágenes de las personas que colaboraron para hacer gestos faciales. Se utilizaron 952 imágenes faciales para las pruebas de detección de las expresiones faciales correspondiente al 50% de las imágenes almacenadas en la base de datos.

En la primera prueba con las imágenes sin normalizar se utilizaron 450 imágenes con el estado emocional de contento, 200 imágenes con el estado emocional de enojado, 152 imágenes con el estado emocional de neutro y 150 imágenes con el estado emocional de sorpresa. Con la utilización de estas imágenes sin aplicar técnica de normalización se obtuvo un resultado de 80.98% de efectividad en el reconocimiento de las expresiones faciales.

Continuando con la segunda prueba utilizando la técnica de normalización de ecualización de histogramas, se utilizaron 380 imágenes con el estado emocional de contento, 250 imágenes con el estado emocional de enojado, 172 imágenes con el estado emocional de

neutro y 150 imágenes con el estado emocional de sorpresa. Con la utilización de estas imágenes aplicando la técnica de normalización de ecualización de histogramas se obtuvo un resultado de 89.28% de efectividad en el reconocimiento de las expresiones faciales.

Finalmente, en la tercera prueba realizada aplicando la técnica de Tan y Triggs se utilizaron 470 imágenes con el estado emocional de contento, 210 imágenes con el estado emocional de enojado, 160 imágenes con el estado emocional de neutro y 112 imágenes con el estado emocional de sorpresa. Con la utilización de estas imágenes aplicando la técnica de normalización de Tan y Triggs, obtuvo un resultado de 92.33% de efectividad en el reconocimiento de las expresiones faciales [42].

Capítulo 5 Pruebas y Resultados

En este trabajo se presenta una nueva arquitectura utilizada para clasificar algunos estados emocionales básicos. La arquitectura propuesta se basa en la utilización de dos técnicas que han demostrado ser eficientes cada una en su propósito, éstas se refieren a los histogramas de gradientes orientados (HOG) y a las máquinas de vectores soporte (MVS). Mientras que los HOG han sido muy utilizados como técnica para extraer y representar características obtenidas de imágenes digitales, las MVS han demostrado ser una técnica de clasificación eficiente que proporciona resultados aceptables aunque se cuente con pocos datos para el entrenamiento.

Las pruebas que se van a presentar tienen que ver con el impacto que la etapa de preprocesamiento causa en la etapa final del proceso (clasificación). En trabajos similares no se ha presentado el impacto que tiene esta etapa en el rendimiento del sistema y se centran más en la evaluación de diferentes clasificadores. Ya que el aspecto medular del trabajo consiste en el análisis de características del rostro relacionadas con la expresión facial, un aspecto importante en este proceso tiene que ver con la etapa de mejoramiento de la imagen de entrada, específicamente lo relacionado con la normalización.

Por lo tanto, se considera que es de interés evaluar el impacto de esta tarea en el rendimiento del clasificador. En este sentido se van a presentar las siguientes pruebas:

1).- Utilización de imágenes sin normalizar.

2).- Normalización de imágenes utilizando la técnica de HOG [15].

3).- Normalización de imágenes aplicando la técnica propuesta por Tan y Triggs [77].

Las imágenes utilizadas fueron recopiladas con el consentimiento y colaboración de estudiantes del Instituto tecnológico de Acapulco (ITA). En la figura 5.1 se muestran algunas de ellas. Las imágenes se obtuvieron en condiciones reales de iluminación y en diferentes escenarios. La resolución inicial de las mismas es de 640 x 480 píxeles. El tamaño inicial fue reducido posteriormente con fines de eficientizar el proceso y una vez que se seleccionó de la imagen solamente el rostro a un tamaño de 80 x 96 píxeles. Las imágenes fueron capturadas utilizando el modelo de color estándar de 3 paletas (RGB), posteriormente y también con fines de eficientizar el proceso se convirtieron a imágenes en escala de gris. Se almacenaron y utilizaron un total de 1904 imágenes correspondientes a la expresión de cuatro emociones básicas (contento, enojado, neutro y sorpresa), fueron utilizadas el 50% de estas para el entrenamiento de las MVS´s y el 50% para evaluarlas. Los resultados obtenidos en estas pruebas se presentan a continuación:

5.1 Prueba 1. Clasificación de Emociones Básicas sin Normalización de las Imágenes Utilizadas.

La normalización de imágenes digitales es un proceso que consiste en homogeneizar los niveles de gris de la imagen original con la finalidad de ampliar su contraste y por ende resaltar algunas características para su uso posterior. En esta prueba se van a utilizar imágenes sin normalizar sus niveles de intensidad, tal como se muestra en la figura 5.2.

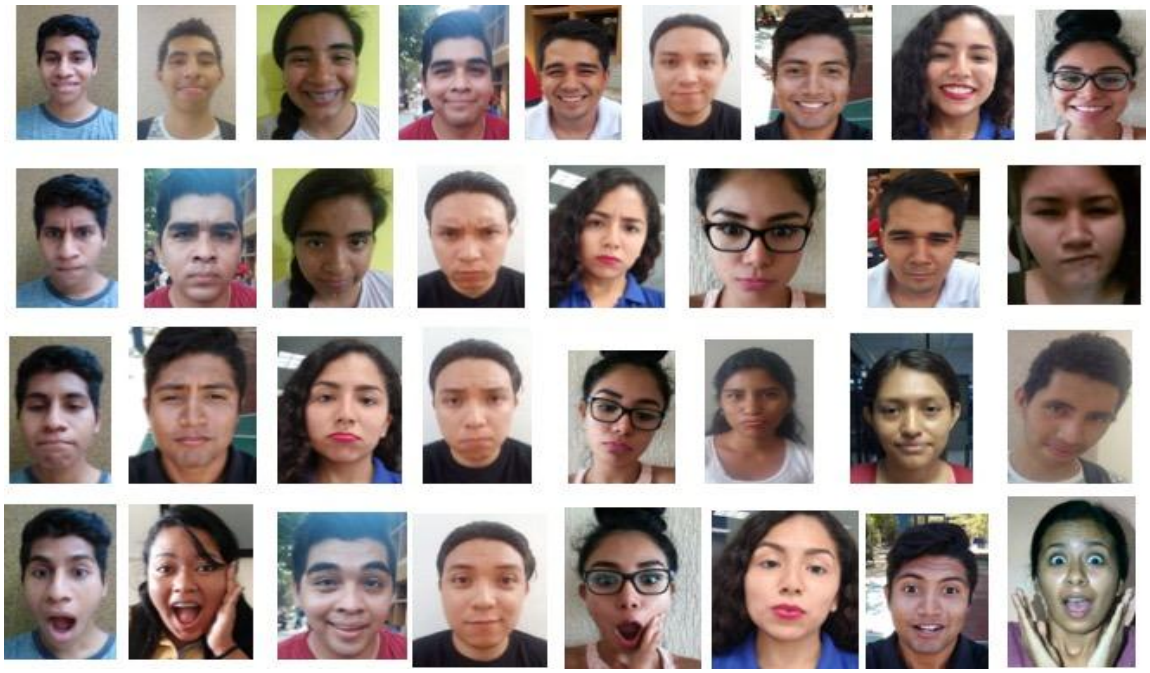
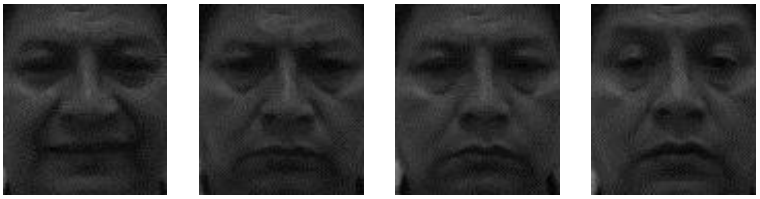


Figura 5.1. Muestra de algunas imágenes mostrando estados emocionales básicos con la participación de alumnos del ITA.

Se analizaron un total de 1904 rostros mostrando cuatro estados emocionales diferentes cada uno. Las características representadas a través de un vector numérico (ver Sección 4.6) se utilizaron para entrenar clasificadores basados en MVS´s. Participaron 476 personas mostrando cuatro diferentes estados emocionales. Se utilizaron 952 estados emocionales para la etapa de entrenamiento, y la misma cantidad de imágenes se utilizaron para la evaluación del modelo. Las imágenes fueron seleccionadas de forma aleatoria. La evaluación del modelo arrojó los datos mostrados en la Tabla 5.1.



a). Contento b). Enojado c). Neutro d). Sorpresa

Figura 5.2. En esta imagen se muestran los cuatro estados emocionales básicos sin normalizar.

Tabla 5.1. Matriz de confusión mostrando las cantidades de estados emocionales que fueron reconocidos de forma exitosa y los que fueron confundidos con otros estados.

	Contento	Enojado	Neutral	Sorpresa
Contento	386	8	22	60
Enojado	18	382	64	12
Neutral	16	58	378	24
Sorpresa	20	34	26	396

A través de estos resultados se desprende que la confiabilidad del modelo para detectar estos cuatro estados emocionales básicos es del 81%, mientras que el porcentaje de clasificación de error es del 19%.

5.2 Prueba 2. Utilización de Imágenes Normalizadas Aplicando la Ecualización del Histograma.

En la segunda prueba, las imágenes utilizadas tanto para el entrenamiento del modelo como para la evaluación del mismo fueron normalizadas aplicando la técnica de ecualización del histograma, como ya sabemos el efecto de esta técnica trae como efecto que las imágenes presenten un mayor contraste ocasionando que ciertas características se muestren resaltadas (figura 5.3) sin perder la estructura de la misma. Al igual que en la prueba anterior se utilizaron un total de 952 imágenes, con 476 representando los diferentes estados emocionales que se pretenden identificar. Se utilizó el 50% de las imágenes con los estados emocionales detectados, estas imágenes fueron seleccionadas de forma aleatoria, para entrenar al modelo y el 50% restante de las imágenes se utilizó para la evaluación. Los resultados obtenidos en las pruebas se muestran en la Tabla 5.2.

a). Contento b). Enojado c). Neutro d). Sorpresa

Figura 5.3. Ilustración de los cuatro estados emocionales básicos donde se ha aplicado la ecualización del histograma.

Tabla 5.2. Matriz de confusión mostrando los resultados de la clasificación de cuatro estados emocionales básicos al utilizar imágenes normalizadas aplicando la ecualización del histograma.

	Contento	Enojado	Neutral	Sorpresa
Contento	424	6	26	20
Enojado	10	420	38	8
Neutral	14	26	426	10
Sorpresa	34	4	8	430

En este caso se observa que la confiabilidad del modelo para clasificar los estados emocionales mostró un incremento en relación con la prueba anterior al pasar de 81% a 89%. Este incremento se justifica debido a que las imágenes fueron resaltadas en algunas características, tonalidad de la intensidad, mientras otras se mantenían sin cambio, bordes.

5.3 Prueba 3. Imágenes Normalizadas Aplicando la Técnica Propuesta por Tan y Triggs.

La tercera prueba que se realizó consistió en clasificar las emociones básicas aplicando la técnica de normalización Tan y Triggs en el preprocesamiento, a las imágenes de entrada (figura 5.4). Este método consiste en aplicar filtros en el dominio del espacio a las imágenes de entrada con la finalidad de resaltar los bordes de la misma (figura 5.4). La intención en este caso consiste en resaltar líneas y bordes o gradientes que identifiquen de forma específica cada emoción básica. Como en la etapa de extracción de las características se aplica la técnica de histogramas de gradientes orientados, esta fase del procesamiento ofrece una ventaja a las dos anteriores.

a).-Contento b).-Enojado c).-Neutro d).-sorpresa

Figura 5.4. En esta imagen se muestran los cuatro estados emocionales básicos aplicando la técnica de normalización Tan y Triggs [77].

En la Tabla 5.3 se muestran los resultados de clasificar cuatro estados emocionales básicos aplicando esta técnica de normalización. En este caso la confiabilidad del clasificador es del 92% mostrando un error del 8%, el incremento de la confiabilidad se debe a que las características que interesan para formar el vector utilizado en las etapas de entrenamiento y evaluación fueron resaltadas en la normalización.

Tabla 5.3 Matriz de confusión donde se ilustran resultados para clasificar cuatro estados emocionales básicos al aplicar la técnica de Tan y Triggs en la normalización de las imágenes de entrada [77].

	Contento	Enojado	Neutral	Sorpresa
Contento	442	6	2	26
Enojado	6	434	24	12
Neutral	10	18	446	2
Sorpresa	24	2	14	436

Equipo utilizado. La implementación y uso del sistema diseñado para la identificación automática de estados emocionales básicos a través del análisis facial requiere lo siguiente:

1).- Una computadora laptop o de escritorio con las siguientes características mínimas:

a).- Disco duro de 500 Gb

b).- 8 Gb de memoria en RAM

c).- Procesador de 3.9 Ghz. de Vel.

d).- Cámara web instalada de 15 Mg pixeles

2. Software:

a).- Windows 8 o superior.

b).- Lenguaje C++, en caso de ser necesario

c).- Office 2010 o superior

Capítulo 6 Conclusiones y Trabajo Futuro

El reconocimiento de emociones humanas de manera automatizada es actualmente un campo activo de investigación debido a su amplia variedad de aplicaciones. Las emociones humanas a las que nos enfocamos en este trabajo son aquellas identificadas como "emociones positivas" (neutral, felicidad) por los psicólogos, ya que cuando una persona se encuentra en este estado emocional resulta más eficiente al afrontar diversas situaciones, tal como aprender en una sesión de enseñanza-aprendizaje. En este sentido, resulta de gran interés el reconocimiento de estas emociones bajo condiciones normales, es decir, que la persona se encuentre en escenarios no controlados y su comportamiento sea normal, no actuado.

Aunque se han logrado avances significativos en la identificación automática de las emociones humanas, éste sigue siendo un problema no resuelto, por lo tanto, cualquier aportación aún siendo modesta resulta de gran utilidad.

En este trabajo se mostró una arquitectura basada en histogramas de gradientes orientados y máquinas de vectores soporte para realizar esta tarea. Aunque estas técnicas ya se han utilizado de forma conjunta en otras aplicaciones (reconocimiento de personas, seguimiento de objetos) no se habían enfocado al reconocimiento del estado emocional humano. Así mismo se mostró el impacto que tiene la selección de técnicas apropiadas durante la etapa de visión de nivel bajo (mejoramiento de la imagen) en las fases posteriores. Cabe mencionar que aunque el tema de seleccionas la técnica apropiada en la fase de mejoramiento impacta en las etapas posteriores, pocas veces se han mostrado estos resultados de forma explícita.

Fueron utilizadas un total de 952 imágenes. Es decir, participaron 476 personas, la mayoría estudiantes del Instituto Tecnológico de Acapulco, a cada uno de ellos se les pidió que actuaran de forma natural al mostrar sus emociones. Asimismo, las imágenes fueron capturadas en condiciones ambientales normales, tal como se muestra en la imagen 5.1.

Cabe mencionar que el número de imágenes utilizadas no es comparable al "*corpus*" utilizado por otros trabajos que utilizan decenas de miles. Esta cantidad de imágenes impacta en la confiabilidad del modelo obtenido, ya que la mayoría de las representaciones obtenidas son "data-driven", es decir, modelos dirigidos por datos, esto significa que entre mayor sea la cantidad de datos utilizados en la fase de entrenamiento, mayor es la confiabilidad del mismo en la fase de reconocimiento.

Sin embargo, existen técnicas que generan buenos resultados aún con cantidades limitadas de datos, tal como las máquinas de vectores soporte (MVS), las cuales son una técnica que ha sido utilizada bajo estas condiciones y obtienen resultados satisfactorios.

En este trabajo se utilizó una cantidad limitada de imágenes representando las emociones humanas de interés.

Se propuso integrar la confiabilidad de las MVS´s debido a su capacidad de generar buenos resultados aún cuando trabaja con datos limitados. Asimismo, se utilizó la técnica de HOG´s debido a su robustez ante condiciones variantes de iluminación, escalamiento y rotación de las imágenes. Estas condiciones son las encontradas en las imágenes cuando se obtienen ante escenarios no controlados.

Asimismo, se pensó en probar diferentes técnicas de mejoramiento de las imágenes obtenidas para decidir cuál ofrecía mejores resultados.

Por lo tanto, podemos concluir lo siguiente:

I. La arquitectura planteada inicialmente, basada en utilizar la técnica de HOG para la extracción de características integrada con las MVS´s como mecanismo de clasificación genera resultados satisfactorios cuando se trabaja con imágenes obtenidas bajo condiciones no controladas.

II. Las técnicas de filtrado que tienen como objetivo resaltar los bordes son adecuadas cuando se integran con los HOG como una técnica posterior para extraer características, ya que ambas operan sobre los puntos de mayor intensidad de las imágenes.

III. Los clasificadores basados en MVS´s realmente obtienen resultados satisfactorios cuando trabajan con cantidades limitadas de datos.

Asimismo, planteamos las siguientes actividades como trabajos futuros.

I. Para que esta herramienta pueda ser utilizada en la solución de problemas reales requiere una mayor confiabilidad en su capacidad de reconocimiento, por lo que hay que integrar más imágenes para su entrenamiento. De igual manera creemos que en la necesidad de ampliar el abanico en las edades de los participantes.

II. Es conveniente integrar un mayor número de emociones para que esta herramienta tenga un mayor número de aplicaciones. El objetivo de esta herramienta es integrarla en un sistema tutor inteligente con el fin de que ofrezca una tutoría cuando el estudiante se encuentre en condiciones de aprender. Sin embargo, con el fin de no limitar su utilidad es necesario integrar más emociones.

III. Con el fin de darle mayor robustez al reconocimiento es necesario integrar otros aspectos en este proceso, tal como considerar la voz como una variable de interés, también es conveniente integrar otras partes del cuerpo en este mecanismo, tal como las manos u otras variantes.

Por el momento consideramos que la metodología manejada y los resultados obtenidos demuestran que estas técnicas son una alternativa satisfactoria para aplicarse en estas tareas. Asimismo, consideramos que los objetivos inicialmente planteados fueron alcanzados.

Bibliografía

[1] Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological bulletin*, 256-274.

[2] Arranz-Aranda, F., Liu Yin, Q., & López-Camara, J. M. (2011). *Interaccion persona-computador basada en el reconcoieminto visual de manos.* Madrid, España.: Tesis Universidad Complutense de Madrid.

[3] Belongie, S., Malik, J., & Puzicha, J. (2001). Matching shapes. *Computer Vision, Proceedings. Eighth IEEE International Conference on* (págs. 454-461). Vancouver, Canada: IEEE.

[4] Bolme, D. S., Draper, B. A., & Beveridge, J. R. (2009). Average of synthetic exact filters. *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2105-2112.

[5]Bovenkerk, J., Fassbender, S., Feist, F., & Mayrhofer, E. (2006). *Strategies for enhanced Pedestrian and Cylcist friendly Design.* Springer.

[6] Brazey, D., & Gout, C. (2014). An algorithm for automatic people detection from depth map sequences. *European Worshop on Visual Information Processing (EUVIP), 2014 5th European Workshop on* (págs. 1-6). IEEE.

[7] Brunelli, R., & Poggiot, T. (1997). Template matching: Matched spatial filters and beyond. *Pattern recognition* (págs. 751-768). Elsevier.

[8] Castrillón-Santana, M., Lorenzo-Navarro, J., Déniz-Suárez, O., Isern-González, J., & Falcón-Martel, A. (2005). Multiple face detection at different resolutions for perceptual user interfaces. *Conferencia Ibérica sobre Reconocimiento de Patrones y Análisis de Imágenes* (págs. 445-452). Heiderberg, Berlin.: Springer.

[9] Cohn, J. F. (2006). Foundations of human computing: Facial expresion and emotion. In Proceedings of the 8th international conference on Multimodal interfaces (pp. 233-238). ACM. *Actas de la 8va conferencia internacional sobre interfaces multimodales* (págs. 233-238). ACM.

[10] Cortés Sanabria, D. F. (2013). *Implementación de un Algoritmo para el Reconocimiento y Análisis de Peatones utilizando Visión por Computador.* Bogota D. C., Colombia: Facultad de Ingeniería pontificia Universidad Javeriana.

[11] Cortes, C., & Vapnik, V. (1995). Support Vector Networks. Machine learning. (págs. 273-297). Springer.

[12] Cowie, R., Douglas-Cowie,, E., Karpouzis, K., Caridakis, G., Wallace, M., & Kollias, S. (2008). *Recognition of emotional states in natural human-computer interaction. In Multimodal user interfaces (pp. 119-153). Springer, Berlin, Heidelberg.* Springer.

[13] D´ Mello, S., & Calvo, R. (2013). Beyon the basic emotions: what should affective computing compute. *CHI'13 Resúmenes Extendidos sobre Factores Humanos en Sistemas de Computación* (págs. 2287-2294). ACM.

[14] Dalal, N. (2006). *Finding People in Images and Videos.* Institut National Polytechnique de Grenoble-INPG.

[15] Dalal, N., & Triggs, B. (2005). Histograms of Oriented Gradients for Human Detection, Research Proyect. *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on* (págs. 886-893). Montbonnot, France.: IEEE.

[16] Dalal, N., & Triggs, B. (2005.). *Histograms of oriented gradients for human detection. In Computer Vision and Pattern Recognition, 2005.* IEEE Computer Society.

[17] Dang, L., Bui, B., Vo, P. D., Tran, T. N., & Le, B. H. (2011). Improved HOG Descriptors. *Knowledge and System Engineering (KSE)*, 186-189.

[18] Darwin, C. (1904). *The Expression of the Emotions in Man and Animals.* AK Peters/CRC Press.

[19] Ekman, P. (1994). Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique. *Psycological Bulleting*, 268-287.

[20] Ekman, P., & Friesen, W. (1978). *Facial action coding system (FACS): manual.* Consulting Psychologists Press Palo Alto.

[21] Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Revista de personalidad y psicología social*, 124-129.

[22] Ekman, P., & Friesen, W. V. (1977). *Manual para el sistema de codificación de acción facial.* Consulting Psychologsts Press.

[23] Ekman, P., & Rosenberg, E. L. (1997). *What the face reveals: basic and applied studies of spontaneous expression using the facial action coding system (FACS).* Oxford: Oxford University Press, USA.

[24] Ekman, P., Friesen, W. V., & Ellsworth, P. (2013). *Emotion in the human face: Guidelines for research and integration of finding.* Elsevier.

[25] Ekman, P., Friesen, W. V., & Hager, J. C. (2002). *Facial Action Coding System Investigator's Guide, A human face.* Salt Lake City, UT.: Salt Lake City UT.

[26] Enzweiler, M., & Gavrila, D. (2009). Monocular pedestrian detection: survey and experiments. *IEEE Transactions on Pattern Analysis & Machine Intelligence* (págs. 2179-2195). IEEE.

[27] Fasel, B., & Luettin, J. (2003). Automatic facial expression analysis: a survey. Pattern recognition. *Reconocimiento de patrones* (págs. 259-275). Elsevier.

[28] Felzenswalb, P., MacAllester, D., & Ramanan, D. (2008). *A discriminatively trained, multiscale, deformable part model. In Computer Vision and Pattern Recognition,.* CVPR 2008. IEEE Conference.

[29] Felzenzwalb, P. F., Girshik, R. B., McAllester, D., & Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 1627-1645.

[30] Fernández, D. (2008). *Stereo vision based pedestrian detection system for assisted driving.* Alcalá, España.: Universidad de Alcalá.

[31] Filko, D., & Martinovic, G. (2013). *Filko, D., Martinovic, G.: Emotion recognition system by a neural network based facial expression analysis. AUTOMATIKA 54(2), 263-272 (2013).* Taylor & Francis.

[32] Flores, M. J., Robayo, D. J., & Saa, D. A. (2015). Histograma del gradiente con múltiples orientaciones (hog-mo) detección de personas. *Revista Vinculos*, 138-147.

[33] Freeman, W. T., & Roth, M. (1995). Orientation histograms for hand gesture recognition. *International workshop on automatic face and gesture recognition* (págs. 296-301). Zurich, Switzerland: IEEE Computer Society .

[34] Freeman, W. T., Tanaka, K., Ohta, J., & Kyuma, K. (1996). Computer vision for computer games. *2nd International Conference on Automatic Face and Gesture Recognition,* (págs. 100-105). Killington, VT, USA,: IEEE.

[35] Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 119-139.

[36] Gandhi, T., & Travedi, M. (2007). Pedestrian protection Systems: Issuess, survey and challenges. *IEEE Transactions on intelligent Transportation systems* (págs. 413-430). IEEE.

[37] Gerónimo Gómez, D. (2010). *A global approach to vision-based pedestrian detection for advanced driver assistance systems.* Barcelona, España: Universitat Autónoma de Barcelona.

[38] Gosavi, A. P., & Khot, S. R. (2013). Facial expression recognition using principal component analysis. *Int. Journals of Soft Computing and Engineering* (págs. 258-262). IEEE.

[39] Hester, C. F., & Casasent, D. (1980). Multivariant technique for multiclass pattern recognition. *Applied Optics* (págs. 1758-1761). Optical Society of America.

[40] Hilario, C. (2008). *Detección de peatones en el Espectro Visible e Infrarrojo para un Sistema Avanzado de Asistencia a la Conducción.* Madrid, España.: Universidad Carlos III de Madrid.

[41] Jiménez, L., & Rengifo, P. (2010). Al Interior de una Máquina de Soporte Vectorial. *Revista de Ciencias*, 73-85.

[42] Jiménez-Váquez, M., Montero-Valverde, J. A., Martínez-Arroyo, M., & Carranza-Gómez, J. (2018). Reconocer emociones básicas a través del análisis facial. *Memorias Academia Journals*, (Autorización para impresión).

[43] Jiménez-Vázquez, M., Montero-Valverde, J. A., Martínez-Arroyo, M., & Carranza-Gómez, J. (2017). Alineación facial para reconocer emociones humanas. *Memorias Academia journals*, 3344-3349.

[44] Keller, C., Enzweiler, M., & Gavrila, D. (2011). A New Benchmark for Stereo-Based Pedestrian Detection. *Intelligent Vehicles Symposium (IV), 2011 IEEE* (págs. 691-696). IEEE.

[45] Kelly, P., O´Connor, N. E., & Smeaton, A. F. (2008). A framework for evaluating stereo-based pedestrian detection techniques. *IEEE Transactions on Circuits and Systems for Video Technology* (págs. 1163-1167). IEEE.

[46] Kotsia, I., & Pitas, I. (2007). Facial Expression Recognition in Image Sequences Using Geometric Deformation Features and Support Vector Machines. *IEEE transactions on image processing* (págs. 172-187). IEEE.

[47] Lopez, J. M., Cearreta, I., Garay, N., Lopez de Ipiña, K., & Beristain, A. (2006). Creación de una base de datos emocional bilingüe y multimodal. *Proc. of the 7th Spanish Human Computer Interaction Confeerence* (págs. 55-66). Proceeding of the 7th Spanish Human Computer Interaction Conference, Interacción.

[48] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision* (págs. 91-110). Springer.

[49] Lyons, M., Akamatsu, S., Kamachi, M., & Gyoba, J. (1998). Coding facial expressions with gabor wavelets. *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, 200-205.

[50] Ma, Y., Ding, X., Wang, Z., & Wang, N. (2004). Robust precise eye location under probabilistic framework. *Null*, 339.

[51] Mahalanobis, A., Viljaya-Kumar, B. V., & Casasent, D. (1987). Minimum average correlation energy filters. *Applied Optics* (págs. 3633-3640). Optical Society of America.

[52] Mallick, S. (6 de December de 2016). *www.learnopencv.com.* Obtenido de www.learnopencv.com/histogram-of-oriented-gradients: http://www.learnopencv.com/histogram-of-oriented-gradients

[53] Min, K., Son, H., Choe, Y., & KIm, Y. G. (2013). Real time Pedestrian Detection Based on a Hierarchical Two-Stage Support Vector Machine. *Industrial Electronics and Applications (ICIEA), 2013 8th IEEE Conference on* (págs. 114-119). IEEE.

[54] Mohan, A., Papageorgiou, C., & Poggio, T. (2001). Example-based object detection in images by components. *Transacciones IEEE en el análisis de patrones \ & Machine Intelligence*, 349-361.

[55] Montagu, J. (1994). *The Expression of the Passions: The Origin and Influence of Charles Le Brun's 'Conférence sur l'expression générale et particuliere.* Yale University Press.

[56] Munder, S., & Gavrila, D. (2006). An experimental study on pedestrian classification. *IEEE transactions on pattern analysis and machine intelligence*, 1863-1868.

[57] Osorio-Arroyave, E. (2015). *Detección Automatizada de Objetos en Secuencia de Video utilizando histogramas de gradientes orientados.* Pereira, Colombia.

[58] Pandzic, I. S., & Forcheimer, R. (2003). *MPEG-4 facial animation: the standerd, implementation and applications.* John Willey & Sons.

[59] Pantic, M., & Patras, I. (2005). Detecting facial actions and their temporal segments in nearly frontal-view face image sequences. *Systems, Man and Cybernetics, 2005 IEEE International Conference on* (págs. 3358-3363). IEEE.

[60] Pantic, M., & Rothkrantz, L. (2000). Pantic, M., & Rothkrantz, L. J. M. (2000). Automatic analysis of facial expressions: The state of the art. IEEE Transactions on pattern analysis and machine intelligence, 22(12), 1424-1445. *IEEE Transactions on pattern analysis and machine intelligence* (págs. 1424-1445). IEEE.

[61] Pantic, M., Penland, A., Nijholt, A., & Huang, T. (2007). Pantic, M., Pentland, A., Nijholt, A., & Huang, T. S. (2007). Human computing and machine understanding of human behavior: a survey. In Artifical Intelligence for Human Computing (pp. 47-71). Springer, Berlin, Heidelberg. *Artifical Intelligence for Human Computing*, 47-71.

[62] Pantic, M., Pentland, A., Nijholt, A., & Huang, T. S. (2006). Human computing and machine understanding of human behavior: a survey. *Artifical Intelligence for Human Computing* (págs. 239-248). Berlin, Heiderberg.: Springer.

[63] Papageorgiou, C., & Poggio, T. (2000). A trainable system for object detection. *International journal of computer vision*, 15-33.

[64] Perez-Gaspar, L. A., Caballero-Morales, S. O., & Trujillo-Romero, F. (2015). Pérez-Gaspar, L. A., Morales, S. O. C., & Trujillo-Romero, F. (2015). Factores en el reconocimiento facial de emociones y la integración de optimización evolutiva. Research in Computing Science, 91, 45-56. *Research in Computing Science*, 45-56.

[65] Perez-Gaspar, L., Caballero-Morales, S. O., & Trujillo-Romero, F. (2014). Pérez-Gaspar, L. A., Morales, S. O. C., & Trujillo-Romero, F. D. J. (2014). Error Modelling Approach based on Artificial Neural Networks for Face Emotion Recognition. *Research in Computing Science*, 21-30.

[66] Phillips, P. J., Moon, H., Rizvi, S. A., & Rauss, P. J. (2000). The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on pattern analysis and machine intelligence* (págs. 1090-1104). IEEE.

[67] Picard, R. W. (1997). *{Computación afectiva. Cambridge, Massachustes Institure of Technology.* The MIT Press.

[68] Ramírez, G. A., Baltrusaitis, T., & Morency, L. P. (2011). Modeling latent discrininative dinamiyc of affective signals. In 1st International audio/Visual Emotion Challenge and workshop in conjuntion with Affective Computing and Intelligent Interaction. *Affective Computing and Intelligent Interaction* (págs. 396-406). Springer.

[69] Rao, K. S., Saroj, V. K., Maity, S., & Koolagudi, S. G. (2011). Recognition of emotions from video using neural network models. *Expert Systems with Applications* (págs. 13181-13185). Elsevier.

[70] Refrégier, P. (1991). Optimal trade-off filters for noise robustness, sharpness of the correlation peak, and Horner efficiency. *Optics Letters* (págs. 829-831). Optical Society of America.

[71] Robinson, P., & Kaliouby, R. (2009). Robinson, P., & El Kaliouby, R. (2009). Computation of emotions in man and machines. *royalsocietypublishing*, 3441-3447.

[72] Rodríguez-Fernández, J. M. (2014). *Computer vision for pedesterian detection using histograms of oriented gredient.* Barcelona, España.

[73] Sashua, A., Gdalyahu, A., & Hayun, Y. (2004). Pedestrian detection for driving assistance systems: Single-frame classification and system level performance. *Intelligent Vehicles Symposium, 2004 IEEE* (págs. 1-6). IEEE.

[74] Savvides, M., & Kumar, B. V. (2003). Efficient design of advanced correlation filters for robust distortion-tolerant face recognition. *Advanced Video and Signal Based Surveillance, 2003. Proceedings. IEEE Conference on* (págs. 45-52). IEEE.

[75] Schmid, K. L., & Cohn, J. F. (2001). Human facial expresion as adaptations: Evolutionary questions in facial expression research. Yearbook of Physical Antology. *American Journal of Physical Anthropology:*, 3-24.

[76] Shih, F. Y., Chuang, C. F., & Wang, P. (2008). Shih, F.Y., Chuang, C.-F.,Wang, P.S.P.: Performance comparisons of facial expression recognition in JAFFE database. International Journal of Pattern Recognition and Artificial Intelligence 22(3), 445-459 (2008). *International Journal of Pattern Recognition and Artificial Intelligence* (págs. 445-459). World Scientific.

[77] Tan, X., & Triggs, B. (2010). Enhanced local texture feature sets for face recognition under difficult lighting conditions. IEEE transactions on image processing, 19(6), 1635-1650. *IEEE transactions on image processing*, 1635-1650.

[78] Tetik, Y., & Bolat, B. (2011). Detection of pedestrians from still images. *Signal Processing and Communications Applications (SIU), 2011 IEEE 19th Conference on* (págs. 670-673). IEEE.

[79] Thuseethan, S., & Kuhanesan, S. (2014). Eigenface based recognition of emotion variant faces. *Computer Engineering and Intelligent Systems*, 31-37.

[80] Video, M. (1998). *SNHC "Text of ISO/IEC FDIS 14 496-3: audio".* Mtg.: Atlantic City MPEG Mtg.

[81] Vijaya-Kumar, B. V. (1986). Minimum-variance synthetic discriminant functions. *J. Opt. Soc. American A.* (págs. 1579-1584). Optical Society of America.

[82] Viola, P., & Jones, M. J. (2004). Robusta detección de rostro en tiempo real. *revista Internacional de visión por computador*, 137-154.

[83] Wang, P., Green, M. B., Ji, Q., & Wayman, J. (2005). Wang, P., Green, M. B., Ji, Q., & Wayman, J. (2005, June). Automatic eye detection and its validation. In null (p. 164). IEEE. *Null* (pág. 164). Washintong: IEEE.

[84] Wang, X., Wang, M., & Li, W. (2014). Scene-specific pedestrian detection for static video surveillance. *IEEE transactions on pattern analysis and machine intelligence* (págs. 361-374). IEEE.

[85] Xu, F., & Gao, M. (2010). Human detection and tracking based on HOG and Particle Filter. *Image and Signal Processing (CISP), 2010 3rd International Congress on* (págs. 1503-1507). IEEE.

[86] Yildirim, S., Bulut, M., Lee, C. M., Kazembadeh, A., Busso, C., Deng, C., . . . Narayanan, S. (2004). Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Deng, Z., Lee, S., ... & Busso, C. (2004). An acoustic study of emotions expressed in speech. . *In Eighth International Conference on Spoken Language Processing.*, 2193-2196.

[87] Yu, F., Chang, E., Xu, Y. Q., & Shum, H. Y. (2001). Emotion detection from speech to enrich multimedia content. *In: Proc. IEEE Pacific-Rim Conferencia Multimedia* (págs. 550-557). Springer.

[88] Zhang, Z., Lyons, M., Schuster, M., & Akamatsu, S. (1998). Comparison betwen facial expression recognition based on geometry and gabor wavelets using multilayer perceptron. *In Automatic Face and Gesture Recognition* (págs. 454-459). IEEE.

[89] Zhihong, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A survey of affect recognitions methods: Audio, visual and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence* (págs. 39-58). IEEE.