Recibido: 30 de septiembre del 2019
Aceptado: 18 de noviembre del 2019
doi: 10.26439/interfases2019.n012.4638:

Sistema de reconocimiento de gestos faciales captados a través de cámaras para analizar el nivel de satisfacción de clientes en restaurantes

Edwin Lara-Lévano
20142424@aloe.ulima.edu.pe
Universidad de Lima. Lima, Perú

Resumen

La presente investigación tiene como principal objetivo el desarrollo de un sistema que reconozca la satisfacción o insatisfacción de un cliente en un restaurante con base en los gestos que este mismo realiza al momento de recibir el servicio brindado por el establecimiento. La implementación del sistema cuenta con una serie de etapas comunes al desarrollo de un proyecto de computación visual, las cuales inician con un preprocesamiento de los datos a ser empleados para el entrenamiento del clasificador que se utilizará, en este caso, una máquina de soporte vectorial. Dentro de este preprocesamiento se hace uso del histograma de gradientes, orientados a la detección del rostro dentro de la imagen, para que esta sea recortada solo al contorno de la cara. De esta manera, se continúa con la extracción de los hitos faciales de la imagen, luego se determina la probabilidad de como cada uno de los sentimientos básicos se encuentran presentes en la expresión de la persona y, en función de estas probabilidades, se determina si la persona está satisfecha o no. Se descubrió que el sistema detectaba correctamente la mayoría de las imágenes ingresadas para las pruebas, sin embargo, se dieron algunos casos donde, a pesar de que el cliente se mostraba satisfecho, este producía ciertos gestos de insatisfacción ocasionados por factores externos.

Palabras clave: *satisfacción del cliente, reconocimiento de expresiones faciales, histograma de gradientes orientados, máquina de soporte vectorial, puntos de referencia faciales*

Abstract

Camera-based facial expression recognition system to analyze customer satisfaction level in a restaurant

The main objective of this research is to develop a system that recognizes customers' satisfaction or dissatisfaction in a restaurant based on their facial expressions when receiving the service provided by the establishment. The implementation of the system has a series of stages common to the development of a visual computing project, which will begin with data preprocessing to train the classifier to be used in this case: a support vector machine. This data preprocessing uses the histogram of oriented gradients for detecting a face inside an image, so that only the face outline is cut. In this way, facial landmarks of the image are extracted, the probability that each of the basic feelings appears in the facial expression of people is established, and, based on these probabilities, customers' satisfaction or dissatisfaction is determined. The results show that the system correctly detected most of the images entered for the tests; however, there were some cases in which, despite the fact that customers were satisfied, they showed certain dissatisfaction expressions caused by external factors.

Keywords: *customer satisfaction, facial expression recognition, histogram of oriented gradients, support vector machine, facial landmarks*

1. Introducción

En los últimos años, el uso de cámaras en los establecimientos comerciales, pequeños locales y empresas, públicas o privadas, ha aumentado considerablemente. Su uso, sin embargo, solo se concentra en brindar un servicio de seguridad. Aunque se conoce, por ejemplo, que los videos grabados por las cámaras captan información importante de la persona, entre otras cosas, expresiones faciales. Es de anotar que la expresión facial es uno de los medios que el ser humano tiene para reflejar emociones y deseos; por ello su análisis es un tema en rápido desarrollo y evoluciona gracias a los avances logrados en áreas como la computación visual, el aprendizaje de máquinas y el aprendizaje profundo.

Este artículo de investigación se enfoca en el análisis de expresiones faciales, tomando como sujeto de pruebas a clientes de un establecimiento de comida, con el objetivo de poder determinar si la persona en cuestión está satisfecha o no con el servicio recibido. Para esto se desarrollará un sistema que, mediante una previa comparación con una línea base de expresiones faciales, pueda determinarse si los gestos realizados por el cliente son de satisfacción o insatisfacción en relación a la calidad del servicio prestado por el establecimiento.

En las siguientes secciones se mostrará el estado del arte, donde se presentarán investigaciones pasadas relacionadas con el tema de esta investigación; luego, los antecedentes, donde se analizarán cada una de las técnicas a ser utilizadas para la implementación del sistema. Posteriormente, se explicará la metodología y, finalmente, se desarrollará una discusión de los resultados obtenidos por las pruebas realizadas. Se culminará con una conclusión general sobre la investigación.

2. Estados del arte

La comunicación no verbal es un área que ha estado fuera del ámbito científico-tecnológico durante mucho tiempo y que, en su momento, solo era reconocida por un grupo de profesionales específicos, por ejemplo, los psicólogos o actores. Sin embargo, a inicios y mediados del siglo xx se empezó a tomar real interés en realizar investigaciones sobre cómo se comunica la gente por medio de las expresiones del rostro (Davis, 1994). Uno de los principales investigadores y pioneros es Paul Ekman, quien en uno de sus libros analizó muchos experimentos realizados sobre el rostro desde mediados del siglo pasado, concluyendo que, al ser realizados en conjunto, las expresiones faciales pueden tomarse como indicadores confiables de algunas de las emociones básicas. En 1978 fue publicado el *Sistema de codificación de acciones faciales* (*Facial Action Coding System, FACS*) por Ekman y Friesen, el cual fue ideado como un sistema que mide toda conducta facial visible en cualquier contexto, sin limitarse a las acciones relacionadas con la emoción. De esta manera, se definieron siete emociones básicas: alegría, sorpresa, furia, tristeza, desprecio, disgusto y contento, cada una de ellas con un patrón codificado dentro de las FACS (Ekman y Oster, 1979).

El reconocimiento de expresiones faciales es una de las nuevas tecnologías para comprender qué expresión y qué emoción emite o realiza el ser humano al extraer, analizar y clasificar las características faciales. Para este propósito existen principalmente tres etapas para el análisis y reconocimiento de expresiones, las cuales son: la detección de rostros, la extracción de las características faciales y la clasificación de estas características (Gao, Jia y Jiang, 2015). Describir y detectar el tipo de emoción correcto que una persona siente, a través de los cambios en la apariencia de su rostro, es la cuestión clave dentro de lo que es el reconocimiento de expresiones faciales. Para ello existen dos enfoques que permiten la descripción de las imágenes faciales, el método basado en características geométricas y el basado en características de apariencia (Ryu, Rivera, Kim y Chae, 2017).

El primer método se concentra en los ángulos y las áreas formadas por ciertos puntos de referencia colocados en las imágenes que representan la cara (Acevedo, Negri, Buemi, Fernandez, y Mejail, 2017). En sí, se realiza una codificación entre las relaciones de ubicación de las partes principales que componen el rostro representado en la imagen, como lo son los ojos, la nariz o la boca (Ryu *et al*., 2017). Mientras que el segundo método se enfoca en las características de toda la imagen usando diferentes algoritmos como, por ejemplo, *local binary pattern* (LBP) o *linear discriminat analysis* (LDA).

Dentro de este campo, son muchas las investigaciones que se han realizado sobre el uso de las técnicas de visión computacional para el reconocimiento de objetos, rostros e incluso emociones. Algunas de las aplicaciones expuestas en estudios anteriores son, por ejemplo, el trabajo desarrollado por Lago y Jiménez Guarín (2014), el cual hace referencia al uso de reconocimiento de expresiones, no solo con la intención de identificar las emociones básicas, tristeza, felicidad, ira, sorpresa, disgusto y miedo, sino también tratar de identificar otras emociones no tomadas en cuenta en este tipo de investigaciones, como el aburrimiento, el interés y la confusión. En esta investigación se propuso un modelo de inferencia de las emociones mencionadas. Para lograr la construcción de este modelo, los autores hacen uso del entorno de trabajo LuxandFace SDK que permitió extraer 64 puntos característicos faciales en la cara y a partir de las distancias existentes entre los puntos se pudo obtener una línea de base.

Otro trabajo es el desarrollado por Whitehill, Serpell, Lin, Foster y Movellan en 2014, el cual realiza una grabación de algunos estudiantes interactuando con un *software* de entrenamiento de habilidades cognitivas. El análisis de estos videos se realiza mediante un reconocimiento de cuadro por cuadro en el video en el cual se utiliza una técnica de anotación que permite crear una línea base en cuanto a las imágenes, que servirán de muestra para relacionar el sentimiento de compromiso con el cuadro en el que se encuentra; posteriormente, se pasa a revisar los distintos algoritmos de reconocimiento facial y, finalmente, se hace uso de las técnicas de aprendizaje automático para desarrollar detectores del nivel de compromiso que tiene el alumno.

En cuanto a trabajos realizados con máquinas de soporte vectorial, está el estudio de Xia Li en 2014, en el cual se usa un clasificador para realizar la multiclasificación de las seis emociones básicas (véase, felicidad, sorpresa, miedo, enojo, tristeza y disgusto) en imágenes de rostros de personas.

Uno de los métodos a ser considerados como parte del desarrollo de la propuesta de solución, que será planteada y detallada más adelante, para el problema propuesto, es el *Histogram of oriented gradients* (HOG), el cual fue planteado inicialmente en la investigación desarrollada por Dalal y Triggs en 2005 llamada *Histograms of oriented gradients for human detection*. En esta investigación, se detalla paso a paso como es desarrollado este descriptor que básicamente se resume en el uso de la dirección de los gradientes distribuidos en toda la imagen. Dalal y Triggs desarrollaron una primera implementación de este descriptor, logrando diferenciar a las personas que caminaban por las calles del escenario de fondo y de otros objetos que estaban a su alrededor.

Otra investigación utilizada como referencia es la de Slim, Kachouri y Atitallah (2018), llamada *Customer satisfaction measuring based on the most significant facial emotion*, en la cual los investigadores plantean el desarrollo de un sistema que permite medir si un cliente se encuentra satisfecho, basándose únicamente en tres emociones, que ellos consideran significantes para determinar el estado de satisfacción, felicidad, sorpresa y neutral. Hacen uso de los hitos faciales, detallados posteriormente, para poder determinar los puntos principales en el rostro de la persona, y según la variación de las distancias entre la agrupación de ciertos puntos, establecer qué emoción de las mencionadas, la persona está expresando. Cabe resaltar, además, que la clasificación de estos puntos para determinar la emoción representada se realiza mediante una máquina de soporte vectorial (SVM), empleada para desarrollar la solución propuesta en esta investigación.

3. Antecedentes

3.1 *Facial landmarks*

Uno de los principales enfoques utilizados, tanto para la detección de rostros, como para el reconocimiento de expresiones, es encontrar los *facial landmarks*, los cuales son una serie de puntos en el rostro que toman como referencia ciertas partes de la cara. Las partes de referencia generalmente usadas son los ojos, la boca, la nariz, las cejas, el mentón y los bordes del rostro (Ouanan, Ouanan y Aksasse, 2016).

Principalmente, los puntos de referencia del rostro están organizados en dos tipos de clases: los puntos de referencia principales, aquellos puntos prominentes para la identidad facial (por ejemplo, las esquinas de la boca, los ojos, la punta de la nariz y las cejas); los demás puntos de referencia denominados secundarios (la barbilla, los

contornos de las mejillas, los puntos medios de las cejas y los labios). En los últimos años, han aumentado las investigaciones científicas referentes a la visión artificial con el objetivo de localizar los denominados puntos de referencial faciales usando distintos modelos. Esto se debe a las distintas aplicaciones que realizan la detección de los *facial landmarks*.

Entre las diversas aplicaciones se encuentran la compresión de la expresión, el registro facial, el reconocimiento facial y el seguimiento facial, como también la utilizada para la reconstrucción de modelos de rostros en 3D (Ouanan *et al.*, 2016). Para poder determinar estos puntos característicos del rostro, es necesario el uso de algoritmos que permitan localizarlos. A pesar de las investigaciones realizadas hasta la fecha, la localización de estos puntos es un tema desafiante, considerando entre los principales obstáculos la variabilidad, debido a factores intrínsecos, como por ejemplo la variación de las caras entre los individuos, y a valores extrínsecos, como la oclusión, la iluminación y la resolución de la imagen (Ouanan *et al.*, 2016).

3.2 *Support vector machine*

Las máquinas de soporte vectorial (SVM, por sus siglas en inglés) son una de las principales herramientas establecidas para el área del aprendizaje automático, sobre todo en lo que respecta al trabajo con imágenes (Xia, 2014). Actualmente son utilizadas en una amplia lista de casos, por ejemplo, en el reconocimiento de dígitos escritos a mano, en la categorización de textos e incluso en la identificación de rostros. A diferencia de paradigmas de aprendizaje, como las redes neuronales, en SVM existe una única solución (Campbell y Ying, 2011).

En cuanto a SVM, la mejor manera de dar su explicación es mediante una clasificación binaria. Cabe resaltar que la clasificación, si bien es la función más aplicada de esta técnica, no es para lo único que es utilizada; puede ser requerida, por ejemplo, para realizar una predicción sobre algún tema en particular mediante un análisis de regresión. La SVM, en general, es una máquina de aprendizaje abstracta que aprenderá de un conjunto de datos de entrenamiento y a partir de estos permitirá generalizar y realizar predicciones correctas sobre un grupo de datos de entrada nuevos (Campbell y Ying, 2011).

Con relación a los datos de entrenamiento, estos básicamente son un conjunto de vectores de entrada, generalmente denotados como *Xi*, en el cual cada vector contiene varias características. Estos vectores de entrada están emparejados con sus respectivas etiquetas, denotadas como *Yi*, y a su vez existen *m* pares (i = 1… m) (Campbell y Ying, 2011). Cuando el problema tiene dos clases de datos bien definidas y separadas, el desarrollo del aprendizaje empieza por encontrar un hiperplano de separación que divide la data clasificada para estas dos clases. El hiperplano de separación es generalmente graficado

en representaciones 2D como una línea media que separa un plano, entonces a los dos lados del hiperplano de separación se encontrarán unos puntos de datos etiquetados como yi = +1 y del otro lado estarán etiquetados como yi = –1. Por su parte, el hiperplano de separación se da como w · x + b = 0 (donde · denota el producto interno o escalar), *b* es el sesgo o desplazamiento del hiperplano desde el origen en el espacio de entrada, *x* son puntos ubicados dentro del hiperplano y la normal al hiperplano, los pesos *w*, determinan su orientación (ver figura 1) (Campbell y Ying, 2011).

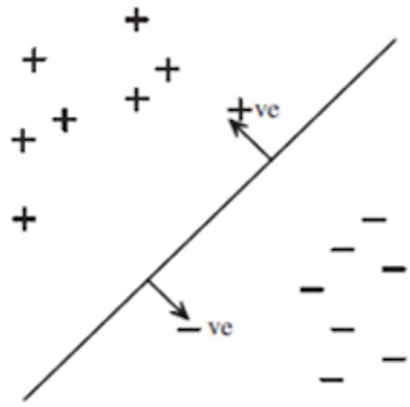


Figura 1. Hiperplano que divide los dos tipos de clases

También está el concepto de hiperplano canónico, formado con los puntos más cercanos de ambas clases (estos puntos son conocidos como vectores de soporte) y la distancia perpendicular entre el hiperplano de separación y un hiperplano canónico es conocido como margen, mientras que la distancia entre los hiperplanos canónicos es denominada como banda de margen (ver figura 4).

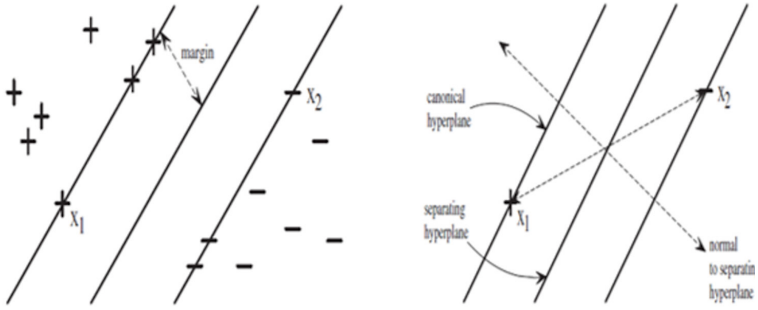


Figura 2. Margen, hiperplano separador e hiperplano canónico

Fuente: Campbell y Ying (2011).

En cuanto a la clasificación binaria mediante SVM se deben considerar dos características importantes:

a) El límite se minimiza maximizando el margen, γ, es decir, la distancia mínima entre el hiperplano que separa las dos clases y los puntos de datos más cercanos al hiperplano.

b) El límite no depende de la dimensionalidad del espacio.

Si se considera una clasificación binaria con puntos de datos Xi, con etiquetas correspondientes Yi = ± 1 y considerando la función de decisión como:

$$f(\boldsymbol{x}) = sign(\boldsymbol{w} \cdot \boldsymbol{x} + b) \qquad (1)$$

donde · es el producto escalar o interno (entonces w · x ≡ wT x).

A partir de esta función de decisión, los datos estarán correctamente clasificados si: yi (w · xi + b)> 0 ∀i desde (w · xi + b) debe ser positivo cuando yi = +1, y debe ser negativo cuando yi = –1.

Para el hiperplano de separación w · x + b = 0, el vector normal es w / ||w||2 (donde ||w||2 es la raíz cuadrada de wT w). Por lo tanto, la distancia entre los dos hiperplanos canónicos es igual a la proyección de x1 − x2 sobre el vector normal w / ||w||2, que da (x1 – x2) · w / ||w||2 = 2 / ||w||2 (ver figura 4, derecha).

Como la mitad de la distancia entre los dos hiperplanos canónicos, el margen es, por lo tanto, γ = 1 / ||w||2. Maximizar el margen es, por lo tanto, equivalente a minimizar:

$$\frac{1}{2}\|w\|_2^2 \qquad (2)$$

sujeto a las restricciones:

$$yi\,(\mathbf{w} \cdot \mathbf{x}i + b) \geq 1 \qquad \forall i \qquad (3)$$

Este es un problema de optimización restringido en el que minimizamos una función objetivo (2) sujeta a las restricciones (3). Como un problema de optimización restringido, la formulación anterior puede reducirse a la minimización de la siguiente función de Lagrange, que consiste en la suma de la función objetivo y las *m* restricciones multiplicadas por sus respectivos multiplicadores de Lagrange. A la siguiente función se le llamará como formulación principal:

$$L(w,b) = \frac{1}{2}(w \cdot w) - \sum_{i=1}^{m} \alpha_i (y_i (w \cdot x_i + b) - 1) \quad (4)$$

donde αi son multiplicadores de Lagrange, y por lo tanto αi ≥ 0. Como mínimo, podemos tomar los derivados con respecto a "b" y "w" y establecerlos en cero:

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{m} \alpha_i y_i = 0 \quad (5)$$

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{m} \alpha_i y_i x_i = 0 \quad (6)$$

Sustituyendo *w* de (6) en *L* (*w*, *b*), obtenemos la formulación dual, también conocida como Wolfe dual:

$$w(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \quad (7)$$

que debe maximizarse con respecto a αi sujeto a las restricciones:

$$\alpha_i \geq 0 \qquad \sum_{i=1}^{m} \alpha_i y_i = 0 \quad (8)$$

Un punto importante a tomar en cuenta es que los datos presentados pueden no ser separables linealmente, con lo cual puede que los datos estén superpuestos en el plano y de esta manera no se pueda hallar un hiperplano capaz de separar las dos clases de datos. Para esto se debe determinar una dimensionalidad mayor a la cual se está trabajando, y además se puede hacer uso de un, por ejemplo, *kernel* gaussiano para poder hallar un hiperplano óptimo. Esta introducción de un *kernel* determinado dentro del modelo se conoce como sustitución de *kernel*.

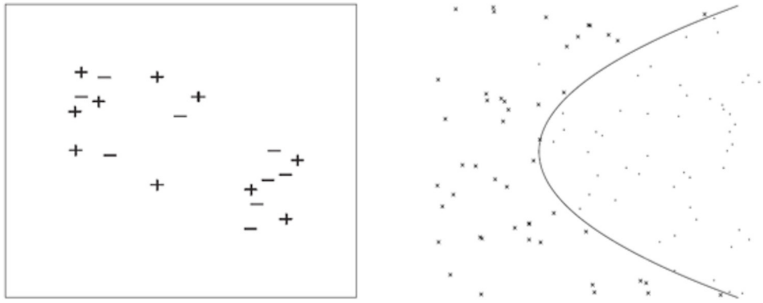


Figura 3. Ejemplo de datos superpuestos en el plano que no pueden ser separados linealmente (izquierda) *vs*. separación de los datos utilizando una nueva dimensión y un *kernel* gaussiano (derecha)

La idea de usar un *kernel* determinado es, como se mencionó, para casos en los cuales el conjunto de datos no pueda ser separado linealmente, por lo cual se hace uso de este tipo de funciones para poder detectar un hiperplano óptimo sobre este conjunto amplio de características. Por definición, una función *kernel* es una función *K*: X x X → R que asigna a cada par de elementos del espacio de entrada, *X*, un valor real correspondiente al producto escalar de las imágenes de dichos elementos en un nuevo espacio *F* (Carmona, 2016).

En cuanto a los diferentes tipos de kernel que existen, algunos ejemplos son:

– *Kernel* lineal:

$$K(x, x') =< x, x' > \tag{9}$$

– *Kernel* polinómico de grado-p:

$$K_p(x, x') = [r < x, x' > + \tau]^p \tag{10}$$

– *Kernel* gaussiano:

$$K(x, x') = \exp\left(-r\left||x - x'|\right|^2\right), r > 0 \tag{11}$$

En cuanto al uso de funciones de *kernel* dentro de lo relacionado al reconocimiento de rostros y expresiones, uno de los más utilizados es el *kernel radial basis function* (RBF), el cual es usado por Xia (2014) para el reconocimiento de expresiones faciales. Este *kernel* es, generalmente, representado de la siguiente manera:

$$K(x, x') = \exp\left(-\frac{||x - x'||^2}{2\sigma^2}\right) \quad (12)$$

En conclusión, la complejidad del uso de una máquina de soporte vectorial (SVM) dependerá solo de la clasificación de los vectores de soporte dentro del plano y así determinar la dimensionalidad que este debe tener para poder indicar correctamente la ubicación del hiperplano de separación.

3.3 Histograma de gradientes orientados

El histograma de gradientes orientados (HOG) es un descriptor de características utilizado para detectar objetos de imágenes, para lo cual este descriptor hace uso de la distribución de direcciones de los gradientes en distintas porciones de una imagen. Este descriptor fue introducido para el reconocimiento de personas por Dalal y Triggs (2005), quienes además revisaron algunas variantes de descriptores HOG, como por ejemplo R-HOG y C-HOG.

En resumen, la idea central de los descriptores HOG es que, tanto la apariencia del objeto como la forma de este, pueden ser descritos mediante el uso de un histograma que represente las direcciones de los bordes de los objetos en la imagen. Para esto, la implementación de este descriptor consta de tres pasos fundamentales (Khan, Rahmani, Ali Shah, y Bennamoun, 2018):

3.3.1 Cálculo del gradiente

El primer paso para el desarrollo del descriptor es el cálculo del gradiente. El gradiente no es más que el cambio direccional en la intensidad de la imagen, la cual es definida por dos valores: la dirección donde se realiza un mayor cambio en cuanto a la intensidad y la magnitud del cambio en esa dirección. En cuanto al cálculo del gradiente, este puede ser calculado de diversas formas, en el caso del descriptor HOG, se realiza con base en la diferencia de intensidad de los pixeles adyacentes en dirección tanto horizontal como vertical. Es así que, si tenemos una imagen I, y se quiere calcular el gradiente en un punto (x, y), se empieza obteniendo la diferencia en la dirección tanto horizontal como vertical, esto es denotado de la siguiente forma:

$$dx = I(x + 1, y) - I(x - 1, y) \quad (1)$$

$$dy = I(x, y + 1) - I(x, y - 1) \quad (2)$$

Donde *dx* y *dy* son la diferencia en la posición horizontal y vertical, respectivamente, de las intensidades de los pixeles. A partir de lo cual se puede calcular la orientación y la magnitud global del gradiente. Para esto se deben tomar en cuenta las diferencias obtenidas anteriormente, y pasarlas a un eje de coordenadas, donde estos dos valores permitirán definir un vector (vector gradiente del pixel), y así calcular la orientación del gradiente en un punto, denotado como θ (*x, y*); además, se puede calcular la magnitud en ese mismo punto, denotado como *g* (*x, y*). Es así como la orientación se definirá como el ángulo que forma el vector con el eje horizontal, el cual puede ser calculado mediante el arco tangente:

$$\theta(x,y) = arctan\frac{dy}{dx} \qquad (3)$$

Mientras que el cálculo de la magnitud del gradiente se determina por la longitud del vector:

$$g(x,y) = \sqrt{dx^2 + dy^2} \qquad (4)$$

Todo lo anterior es a partir del trabajo con imágenes en escala de grises. Si se tomaran en cuenta imágenes a color, el descriptor debe priorizar el color que domina en forma local para cada pixel, con lo cual para cada pixel se debe calcular el gradiente en cada uno de los tres canales de colores existentes (véase RGB, rojo, verde y azul) y luego se toma el gradiente del canal con mayor magnitud (Dalal y Triggs, 2005).

3.3.2 Cálculo del histograma

El segundo paso para el desarrollo del descriptor es el cálculo del histograma de orientaciones para cada celda definida en la imagen. Para lo cual lo primero que debe realizarse es la división de la imagen en un número fijo de celdas de un tamaño determinado. Este tamaño fijo para cada celda, generalmente suele ser entre 6 y 8 pixeles, tanto en alto como en ancho (Khan *et al.*, 2018). Luego, se deberá considerar como se divide el rango de orientaciones en un número de intervalos fijo.

Lo primero a considerar es si la orientación del gradiente se toma con signo o no, para lo cual, si se elige la primera opción, el rango de orientaciones irá desde 0 a 360°; mientras que, si el signo de la orientación no es tomado en cuenta, el rango ira desde 0 a 180°. Con esta última opción, dos gradientes con la misma dirección, pero sentidos inversos se consideran equivalentes y quedan asignados al mismo intervalo. El otro punto a considerar, es en cuántos intervalos se dividirá el rango de orientaciones. Habitualmente, y considerando un rango de 0 a 180°, este rango es

dividido en 9 intervalos, con lo cual cada intervalo agrupa un rango de orientaciones de 20° (Dalal y Triggs, 2005).

3.3.3 Cálculo del descriptor

Uno de los grandes retos que cualquier descriptor de características tiene es el hecho de combatir la invarianza de determinados aspectos en la imagen (véase cambios de escala, iluminación, etc.). En el caso del detector HOG, al ver un cambio de iluminación y contraste en la imagen, la intensidad del gradiente va a cambiar, por lo cual estos cambios de intensidad también se verán reflejados en los histogramas. Para reducir las diferencias en los resultados de estas imágenes, se necesitará normalizar los valores de los histogramas, con el objetivo de que la magnitud global del gradiente sea similar en cada cambio de iluminación (Dalal y Triggs, 2005).

La normalización será realizada en bloques, que no es más que una conjunción de celdas de la imagen. Generalmente el tamaño del bloque es de 2 x 2 o 3 x 3 celdas. Es así que de este bloque se obtendrán los histogramas de cada celda que está contenida dentro de este, los cuales serán agrupados en un vector no normalizado v = (x1, …, xn). Basados en esto, se procede a la normalización del vector que contiene los histogramas concatenados de un bloque, la cual es desarrollada dividiendo el vector sobre una norma. Entonces, considerando ||v||k donde la norma es k para k = 1, 2 y ε es una pequeña constante, se pueden determinar los siguientes factores de normalización:

$$L2-norm: \quad v' = \frac{v}{\sqrt{\|v\|_2^2 + \varepsilon}}, \tag{5}$$

o

$$L1-norm: \quad v' = \frac{v}{\sqrt{\|v\|_1 + \varepsilon}}, \tag{6}$$

o

$$L1-sqrt: \quad v' = \sqrt{\frac{v}{\|v\|_1 + \varepsilon}}, \tag{7}$$

La aparición de la constante ε en las fórmulas tiene como objetivo evitar divisiones por cero en los casos donde la intensidad sea constante en todo el bloque, teniendo como resultado que la magnitud total del gradiente sea cero. Se

debe tomar en cuenta que, la norma de un vector es igual a la raíz cuadrada de la suma de todos sus componentes al cuadrado:

$$\|v\| = \sqrt{\sum_{i=1}^{n} x_i^2} \qquad (8)$$

4. Metodología y experimentación

Antes de comenzar con la explicación de las etapas que se siguieron para el desarrollo de la prueba de concepto, se expondrán las herramientas utilizadas para la implementación de esta. En primer lugar, el desarrollo en general se realizó en una *laptop* con las siguientes características:

- Procesador Intel Core i7-6500 de 2.5 GHz con una RAM de 8GB
- Tarjeta gráfica NVIDIA GEFORCE 940 MX
- Sistema Operativo Windows 10

En cuanto al lenguaje de programación utilizado, la implementación se desarrolló en Python 3.6.2, empleando Miniconda 3. Además, se trabajaron principalmente con las siguientes librerías:

- Scikit-learn: es una Librería para la programación en Python, la cual está especializada en el aprendizaje de máquina.
- NumPy: librería de funciones matemáticas de alto nivel que permite una mejor operación con vectores y matrices.
- Scikit-image: se utiliza básicamente para el procesamiento de imágenes.
- Dlib: librería de *software* multiplataforma escrita en C++. Contiene componentes para trabajar con *machine learning* y procesamiento de imágenes.
- OpenCV: es una librería libre utilizada para los temas de inteligencia artificial.

4.1 Obtención de una base de datos con imágenes de emociones

Como se explicó anteriormente, existe una buena cantidad de bases de datos disponibles para la detección de rostros y el reconocimiento de expresiones. Para esto se tomaron en cuenta dos bases de datos para las respectivas pruebas: la Extended Cohn-Kanade

Dataset (CK+) y la base de datos obtenida de la librería Scikit-learn. La primera fue utilizada para la elaboración del sistema final; mientras que la otra, provisionalmente, para la etapa de preprocesamiento de la imagen. Un ejemplo del contenido de este conjunto de imágenes es mostrado a continuación:

Figura 4. Ejemplo del *dataset* de rostros en Scikit-learn

Figura 5. Ejemplo del *dataset* de rostros en The Extended Cohn-Kanade Dataset (CK+)

Algunos aspectos a tomar en cuenta sobre la base de datos Cohn Kanade es que tiene subdividido cada conjunto de imágenes de emociones según el sujeto de prueba utilizado; además, algunos de los sujetos de prueba no cuentan con una galería de imágenes de todas las emociones que esta base de datos ofrece. Por ejemplo, el sujeto 1 tiene distintas imágenes que muestran las emociones básicas; sin embargo, el sujeto 2 solo tiene imágenes sobre la emoción neutral y la felicidad. Un último punto a tomar en cuenta es que este *dataset* "desarrolla" la creación de la expresión, es decir, cada conjunto de imágenes de emociones representan el paso de un rostro neutral a uno que muestra ciertos gestos que demuestran la emoción. Por lo cual se tuvo que realizar un reordenamiento de las imágenes proporcionadas, separando todas las imágenes en carpetas, cada una según la emoción que se representaba, sin importar el sujeto de prueba que las esté expresando.

Una vez reorganizado el *dataset*, este pasó a un preprocesado para obtener únicamente el área del rostro y desechar el *background* de la imagen. Para ello se usaron los histogramas de gradientes orientados, los cuales recortaron la imagen a solo el área del rostro; además, cada una de estas imágenes fue redimensionada a una escala de 350 x 350 pixeles, esto con el objetivo de estandarizar los tamaños de todas las imágenes obtenidas.

Una vez realizado esto, se pasó a extraer los *facial landmarks* de cada una de las imágenes del *dataset* modificado, los cuales fueron vectorizados para posteriormente servir de entrada para el entrenamiento de la máquina de soporte vectorial usada para la clasificación de las emociones reconocidas en las imágenes. Para ello se empleó 80 % de la data extraída al azar para ser usada como data de entrenamiento, considerando el 20 % restante como data de prueba para el clasificador. Esta máquina contó con un *input* de imágenes que no fueron rostros, con esto se obtuvo un modelo predictor.

4.2 Obtención de videos de personas en un restaurante

Como se mencionó, los videos obtenidos para esta investigación fueron recolectados de la web, siendo debidamente escogidos bajo ciertas características específicas para la finalidad de esta investigación y en favor del desarrollo del sistema. Estas características son, por ejemplo, que los videos deben mostrar la grabación del rostro de la persona a lo largo del servicio recibido en el restaurante; además, para poder validar el resultado del procesamiento, la persona en cuestión deberá dejar su opinión clara y concisa, en algún momento de la grabación, sobre el servicio brindado para así comparar su opinión con los resultados obtenidos. De los seleccionados, se debe tener en cuenta que no todos los minutos del video son necesarios para realizar el análisis propuesto en esta investigación, por lo cual se deberá realizar una extracción de ciertos momentos que puedan ser utilizados por el sistema, es decir, solo fueron necesarios fragmentos en los cuales la persona recibió el servicio y opinó sobre este. Esta extracción de las partes del video la realizó el autor de forma manual.

4.3 Procesamiento de las imágenes captadas del video

Al igual que con las imágenes de las bases de datos obtenidas, en esta etapa se realizó un procesamiento de los *frames* obtenidos del video, el cual comenzó cambiando las imágenes de formato RGB a escala de grises. Una vez que todas estas imágenes fueron unificadas, se usó el histograma de gradientes orientados para realizar el recorte de la imagen solo al área del rostro. Las imágenes fueron pasadas a un tamaño estándar de 350 x 350 pixeles para un uso más rápido del algoritmo, este cambio es de carácter totalmente subjetivo por el autor de este documento y con base en un rango calculado según los trabajos anteriores de otros investigadores.

Figura 6. Ejemplo de características HOG

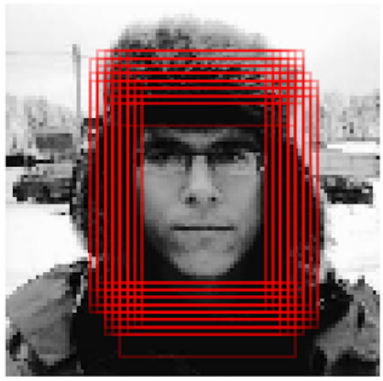


Figura 7. Resultado de la detección del rostro

La imagen presentada es solo una imagen referencial (escogida por el autor para realizar tan solo una prueba inicial sin contexto) de cómo funciona esta parte del sistema. A continuación, se mostrarán ejemplos de imágenes extraídas a partir de los videos obtenidos (figuras 8 y 9) y como fue el resultado del procesamiento final de estas:

Figura 8. Frame original del video del restaurante

Figura 9. Escalas de grises

4.4 Extracción de las características

Es importante tener en cuenta la cantidad de data que será tomada para realizar un análisis de expresiones. Se conoce que un video es un conjunto de imágenes que pasan por segundos de reproducción. Los videos tiene decenas y hasta centenas de *frames* (imágenes) por segundo, la diferencia entre un *frame* y su consiguiente no tendría casi ninguna diferencia, con lo cual el analizarlos traería consigo un gasto computacional innecesario. Por este motivo se decidió obtener un *frame* cada medio segundo del video, considerando este rango de tiempo lo suficientemente amplio para un mejor análisis y diferencia entre imágenes. Los *frames* seleccionados fueron almacenados en una carpeta aparte, la cual sirvió como un nuevo *dataset*, el cual fue utilizado para la extracción de las características (*facial landmarks*).

Una vez las imágenes estuvieron correctamente procesadas, se realizó la extracción de los *facial landmarks*. Para este proceso, se usó la librería Dlib, la cual contiene un predictor de estos puntos característicos sobre el rostro. Cabe recalcar que el predictor no viene junto con la librería, sino que debe ser descargado previamente para su utilización, el nombre de este es "shape_predictor_68_face_landmarks.dat" (ver figura 10).

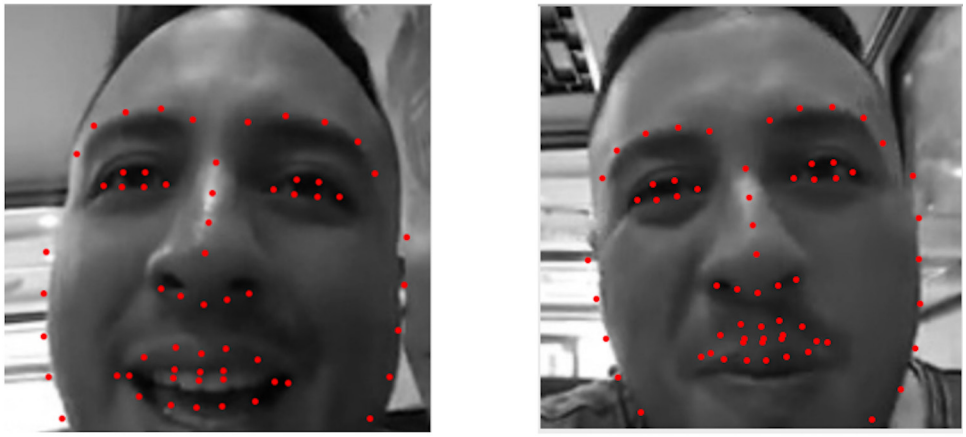


Figura 10. Ejemplo de los *facial landmarks*

Una vez obtenidos estos facial landmarks, las coordenadas correspondientes a cada uno de estos puntos son almacenados. Después se realizó una normalización de las distancias entre los puntos demarcados respecto del centro del rostro. Estos *landmarks* "normalizados", posteriormente, fueron almacenados en un vector mediante el uso de la librería Numpy. El proceso anteriormente detallado se utilizó también con las imágenes obtenidas para el entrenamiento de la máquina de vectores.

4.5 Clasificación de las características

Teniendo el conjunto de datos (*landmarks*) normalizados y agrupados en vectores, estos fueron ingresados a la máquina de soporte vectorial para ser clasificados y así determinar la probabilidad de que cada una de las emociones básicas estén presentes en la imagen analizada. Esta clasificación se basa en el modelo previamente creado con el *dataset* descrito en el primer apartado de esta metodología. Por ejemplo, si tomamos como referencia la imagen del sujeto de prueba de las figuras 6 y 7, luego de todo el proceso descrito en los pasos anteriores, nos daría como resultado el siguiente:

- Enojo: 0,25003043
- Desprecio: 0,02009596
- Disgusto: 0,01237244
- Miedo: 0,0028604
- Felicidad: 0,00137583

- Neutral: 0,67577552
- Tristeza: 0,03366757
- Sorpresa: 0,00382185

Con lo cual el sistema determinaría que la emoción que se expresa en la figura es una emoción neutral, detectando también, de una manera menos clara, el sentimiento de enojo. Considerando que en las investigaciones revisadas sobre este tema no se encontró una manera para determinar un patrón o forma de los gestos de la cara que concluyeran si una persona está satisfecha o no, se decidió que, al momento del servicio brindado, las emociones positivas fueran consideradas como un nivel de satisfacción bueno del cliente; al contrario, las emociones negativas, como insatisfacción del cliente. Es así como, según la investigación realizada por Slim, Kachouri y Atitallah, las emociones que están más relacionadas a la satisfacción serán la felicidad y la sorpresa, mientras que las otras emociones serán consideradas como insatisfacción.

Por ello, se tomaron en cuenta solo las dos emociones más resaltantes destacadas por el SVM para realizar la clasificación de la satisfacción y, además, se consideró que estas emociones debían tener una relevancia en la probabilidad mostrada, para este caso se tomó en cuenta la emoción en caso de que su probabilidad fuese superior a 0,2. Por ejemplo, en los resultados mostrados se puede denotar que la emoción que más probabilidad tiene de estar en la imagen es la neutral, sin embargo, la emoción de enojo es la segunda emoción con mayor probabilidad y, además, su probabilidad es mayor a 0,2, por lo cual el sistema arrojaría como resultado que la persona está insatisfecha. En el caso de tener una probabilidad de emoción que sea superior a 0,2, solo se considerará esta, según su carácter positivo o negativo, para determinar si la persona se encuentra satisfecha o no.

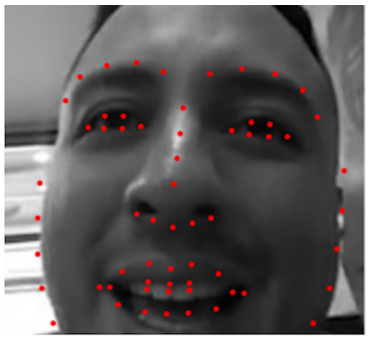
4.6 Evaluación del rendimiento y los resultados obtenidos: resultados del sistema con videos de personas en un restaurante

Para la realización de las pruebas del sistema desarrollado, se tuvieron en cuenta videos obtenidos de la plataforma YouTube de personas que "reaccionaban" a la atención y servicio recibido en distintos restaurantes. Para ejemplificar el desarrollo del sistema, se seleccionó uno de los videos al azar, del cual se obtuvieron 1328 imágenes, teniendo en cuenta que el video duraba alrededor de los 11 minutos y que por cada medio segundo se escogía un *frame* de este. Todas estas imágenes fueron pasadas, primero, por el preprocesamiento mencionado, con lo cual las imágenes quedaron en escala de grises y a 350 x 350 pixeles; además, se eliminó el *background* de cada una, realizando el recorte en el contorno del rostro mediante el uso de los histogramas de gradientes orientados,

sin embargo, no todos los *frames* captados contenían el rostro de una persona, por lo cual solo fueron almacenados y procesados aquellos en los cuales el rostro de la persona sí fue detectado.

Con estos pasos, el conjunto de imágenes a ser analizado disminuyó a 562 imágenes, todas ellas contenían la imagen de un rostro que podía ser analizado. Posteriormente, se pasó a realizar el análisis de gestos mediante la obtención de los *facial landmarks* de cada imagen, con lo cual una por una iba ingresando a la máquina de vectores para la clasificación de las emociones (figura 11).

Ejemplo 1



- Enojo: 0,0249897
- Desprecio: 0,13446804
- Disgusto: 0,02086379
- Miedo: 0,09809341
- Felicidad: 0,01468039
- Neutral: 0,07319924
- Tristeza: 0,115125
- Sorpresa: 0,51858043

Ejemplo 2

- Enojo: 0,0249897
- Desprecio: 0,13446804
- Disgusto: 0,02086379
- Miedo: 0,09809341
- Felicidad: 0,01468039
- Neutral: 0,07319924
- Tristeza: 0,115125
- Sorpresa: 0,51858043

Figura 11. Resultado con los dos rostros mostrados anteriormente

Considerando estos resultados, el sistema muestra que, en el caso del ejemplo 1, la emoción que sobresale es de sorpresa, seguida de la emoción de tristeza, sin embargo, esta última se encuentra por debajo de la probabilidad anteriormente establecida (mayor a 0,2), con lo cual el sistema clasifica que la persona se encuentra satisfecha, puesto que

la única emoción que toma en cuenta para la decisión es la de sorpresa, catalogada como positiva, mientras que la emoción de tristeza es dejada de lado, puesto que no supera el margen establecido para ser considerada como significativa. Por otra parte, en el ejemplo 2, la emoción que más probabilidad tiene es la neutral, sin embargo, la segunda emoción es la tristeza, superando el mínimo establecido para ser considerado por el sistema, con lo cual el sistema clasifica que la persona en esta imagen se encuentra insatisfecha, puesto que la segunda emoción en consideración se encuentra dentro del conjunto de emociones categorizadas como negativas.

Un punto que hasta el momento no se ha abordado en esta investigación es la posibilidad de que dos o más personas aparezcan en la misma escena. Bajo este contexto, el sistema solo tomará en cuenta el rostro de la primera persona que reconozca en la imagen para su análisis. Por ejemplo, se tiene la siguiente imagen, obtenida de un video diferente, en el cual se muestran dos personas en un restaurante. El sistema toma en cuenta solo uno de los rostros y realiza el preprocesamiento anteriormente descrito.

Figura 12. ***Frame*** con dos personas en el restaurante

Figura 13. En escalas de grises

Con estos resultados, el sistema podría sufrir cierta variación debido a que puede estar reconociendo a otra persona que no es el objeto del estudio. Es por ello que, para evitar estas situaciones, los videos a ser analizados deben ser previamente recortados a los tramos en los cuales la persona que esté recibiendo el servicio aparezca sola. Esto no afecta el objetivo final de la investigación, puesto que la finalidad de esta es realizar el análisis individual de una persona y no de manera grupal. Finalmente, todos los *frames* que han sido procesados por el sistema son evaluados de manera conjunta, es decir, que si en conjunto la mayoría de *frames* han sido etiquetados como una imagen donde la persona muestra satisfacción, se considerará que la persona está satisfecha durante todo el proceso del servicio brindado. Esto se realiza debido a que, durante el procesamiento del video extraído, la persona puede mostrar ciertos gestos que pueden ser catalogados como insatisfacción, los cuales no determinan que la persona este insatisfecha con el servicio.

4.7 Validación de los resultados del sistema

En este apartado se debe tomar en cuenta que la forma ideal de validar estos resultados sería contar con la opinión directa del cliente sobre el servicio brindado en el restaurante, mediante el uso de una encuesta de satisfacción de servicio correctamente enfocado. Sin embargo, los videos analizados fueron obtenidos, en su totalidad, mediante la plataforma de internet YouTube, por lo cual realizar este tipo de validación sería imposible. Es por ello que, con el fin de validar los resultados del análisis, los videos elegidos fueron seleccionados bajo ciertos parámetros que permitieron llevar a cabo la validación de los resultados. Por lo tanto, solo se tomaron en cuenta aquellos videos donde la persona que "reacciona" a la atención recibida, posteriormente da su opinión explicita sobre el tema, con lo cual se tendría la forma de relacionar el sentimiento detectado por el sistema (satisfacción o insatisfacción) con lo que verdaderamente opina el cliente, validando de esta manera los resultados obtenidos.

Para ilustrar lo comentado, se tomará en cuenta el ejemplo 2. Según los resultados arrojados por el sistema, la persona fue catalogada como insatisfecha, siendo las emociones con más alta calificación la neutral y la tristeza. Por otro lado, según el video original del cual se sacaron las imágenes, en ese lapso de tiempo el cliente dice explícitamente lo siguiente (basado en los subtítulos generados por la misma plataforma): "Sinceramente, le voy a poner un 5 sobre 10, porque no es como que ¡huy!, normal, no me sorprendió tanto", infiriendo de esto que su expectativa era mayor y que, en cierta medida, no se encontraba contento con lo recibido. De esta manera, se puede afirmar que el sistema clasificó correctamente a la persona como insatisfecha.

5. Discusión

Para realizar una correcta evaluación y prueba del sistema, en general, se tomaron en cuenta diversos videos de varios autores, con distintos rasgos faciales para acopiar la mayor cantidad de casos posibles. Bajo este contexto, los resultados que arrojó el sistema no fueron 100 % correctos, algunos de los casos más importantes a tomar en cuenta se describen a continuación. En primer lugar, el sistema desarrollado no es capaz de determinar el "contexto" de la escena, es decir, que, por ejemplo, si la persona que recibe el servicio en el restaurante llega a este con una actitud negativa, por diversos motivos que se desconocen, posiblemente, aun siendo la atención buena, mostrará expresiones faciales negativas, con lo cual el sistema clasificará que el cliente se muestra insatisfecho, sin embargo, bajo este escenario no se puede determinar que esta insatisfacción sea consecuencia del servicio brindado.

Otro caso a tomar en cuenta es cuando los clientes realizan gestos espontáneos que se encuentran fuera de contexto, por ejemplo, el cliente puede estar recibiendo una buena atención y mostrar gestos de satisfacción por esto, sin embargo, en algún momento de

este proceso (que para el sistema es representado como un *frame*/imagen del momento), el cliente puede realizar algún gesto negativo, debido a algo que le dijo la persona que lo acompaña o por algún motivo externo al servicio, en este caso, el sistema clasifica al *frame* analizado como insatisfacción. Sin embargo, el fin de esta investigación es determinar la satisfacción o insatisfacción del cliente durante todo el proceso del servicio brindado, por lo tanto todos los resultados de cada *frame* utilizado fueron analizados en conjunto.

6. Conclusiones

Las empresas y establecimientos están buscando constantemente acaparar la mayor cantidad del mercado de su rubro en el cual se desenvuelven. Es así como uno de los puntos a ser tomados en cuenta, por no decir el más importante, es el hecho de lograr fidelizar a sus clientes, mejorando los aspectos que hacen que estos se sientan satisfechos con lo que reciben y corrigiendo aquellos que los disgustan. La presente investigación tiene como principal objetivo determinar la satisfacción o insatisfacción de un cliente al momento de recibir un servicio brindado por un restaurante, esto con el fin de poder brindar a este tipo de establecimientos, y en general a todos aquellos en los cuales pueda aplicarse el trabajo desarrollado, una nueva herramienta con la cual se pueda añadir un valor agregado en el proceso de atención al cliente.

Esta investigación hace uso de diversas técnicas y algoritmos de visión computacional como los histogramas de gradientes orientados, los cuales fueron utilizados para la detección del rostro de la persona, el uso de los *facial landmarks*, para determinar los puntos característicos del rostro, con lo cual se identificó qué sentimiento expresaba el cliente según el gesto que mostraba. También se hizo uso de una máquina de soportes vectorial para la clasificación de los datos obtenidos.

En un principio, la presente investigación tuvo como objetivo desarrollar las pruebas del sistema haciendo uso de videos recolectados de manera propia en un restaurante, sin embargo, si esto se realizaba, se debía pedir el consentimiento de cada una de las personas que serían grabadas, limitando la obtención de los videos a la buena voluntad de los clientes. Además, una de las principales características planteadas para esta investigación era obtener gestos naturales expresados por la persona, sin embargo, al ser estas informadas de que serían grabadas, la data obtenida podría verse sesgada. Por ello, se determinó que el proceso de las pruebas del sistema desarrollado se realizaría con imágenes extraídas de videos colgados en la web. Los resultados obtenidos mostraron que la clasificación de la satisfacción e insatisfacción de la persona se realizaron correctamente en la mayoría de los casos; sin embargo, existieron casos, por motivos externos al servicio, y que no se pudieron controlar dentro de esta investigación, en donde el sistema clasificó como insatisfacción ciertos *frames* del video en los cuales no hubo responsabilidad por parte del servicio brindado.

Para posibles trabajos futuros, se deben tomar en cuenta, principalmente, las limitantes que se tuvieron en la realización de esta investigación, como, por ejemplo, el hecho de no poder realizar las pruebas con data propia obtenida de un restaurante. Esta investigación puede ser tomada, además, bajo otro contexto distinto al de clientes en un restaurante, se podría aplicar, por ejemplo, en las ventanillas de un banco o en otro lugar donde el rostro de la persona pueda ser captado directamente. Además, se pueden tomar en cuenta otros tipos de técnicas distintas a las utilizadas en esta investigación para el desarrollo del sistema y, de esa manera, poder tener una comparación en el rendimiento de cada sistema.

Referencias

Acevedo, D., Negri, P., Buemi, M. E., Fernandez, F. G., y Mejail, M. (2017). A simple geometric-based descriptor for facial expression recognition. *12th IEEE International Conference on Automatic Face & Gesture Recognition*, pp. 802-808.

Campbell, C., y Ying, Y. (2011). Learning with support vector machines. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *5*(1), pp. 1-95.

Carmona Suárez, E. J. (2014). Tutorial sobre máquinas de vectores soporte (SVM). Recuperado de http://www.ia.uned.es/~ejcarmona/publicaciones/[2013-Carmona]%20SVM.pdf

Dalal N., y Triggs, B. (2005). Histograms of oriented gradients for human detection. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (CVPR'05), 1, pp. 886-893. doi:10.1109/CVPR.2005.177

Davis, F. (1994). *El lenguaje de los gestos*. Buenos Aires: Emecé Editores.

Ekman P., y Oster, H. (1979). Expresiones faciales de la emoción. *Annual Review of Psychology*, 30, pp. 527-554.

Gao, G., Jia, K., y Jiang, B. (2015). An automatic geometric features extracting approach for facial expression recognition based on corner detection. *International Conference on Intelligent Information Hiding and Multimedia Signal Processing* (IIH-MSP), pp. 302-305. Adelaide, SA, Australia: IEEE. doi:10.1109/IIH-MSP.2015.33

Khan, S., Rahmani, H., Ali Shah, S. A., y Bennamoun, M. (2018). *A guide to convolutional neural networks for computer vision*. Morgan & Claypool Publishers.

Lago, P., y Jiménez Guarín, C. (2013). An affective inference model based on facial expression analysis. *IEEE Latin America Transactions*, *12*(3), pp. 423-429.

Ouanan, H., Ouanan, M., y Aksasse, B. (2016). Facial landmark localization: Past, present and future. *4th IEEE International Colloquium on Information Science and Technology* (CiSt), pp. 487-493.

Ryu, B., Rivera, A. R., Kim, J., y Chae, O. (2017). Local directional ternary pattern for facial expression recognition. *IEEE Transactions on Image Processing*, *26*(12), pp. 6006-6018.

Slim, M., Kachouri, R., y Atitallah, A. (2018). Customer satisfaction measuring based on the most significant facial emotion. *15th IEEE International Multi-Conference on Systems, Signals & Devices (SSD)*, pp. 502-507.

Whitehill, J., Serpell, Z., Lin, Y.-C., Foster, A., y Movellan, J. R. (2014). The faces of engagement: Automatic recognition of student engagementfrom facial expressions. *IEEE Transactions on Affective Computing*, *5*(1), pp. 86-98.

Xia, L. (2014). Facial expression recognition based on SVM. *7th IEEE International Conference on Intelligent Computation Technology and Automation*, pp. 256-259.