DESDE 2013
https://repository.uaeh.edu.mx/revistas/index.php/icbi/issue/archive
Pädi Boletín Científico de Ciencias Básicas e Ingenierías del ICBI

ISSN: 2007-6363

Publicación Semestral Pädi Vol. 10 No. Especial 4 (2022) 110-115

Detección de estados de ánimo en ambientes no restringidos
Detection of mood states in unrestricted environments

M. Sánchez-Ruiz[a], J. Flores-Monroy[a], E. Escamilla-Hernández[a], M. Nakano-Miyatake[a], H. Perez-Meana[a,*]

[a] *Escuela Superior de Ingeniería Mecánica y Eléctrica, Unidad Culhuacan, Instituto Politécnico Nacional, 04440, Ciudad de México, México.*

Resumen

Uno de los factores importantes que influyen en los accidentes automovilísticos es el manejar bajo condiciones no-óptimas, tales como estrés, ira, miedo, depresión entre otros, en las cuales la posibilidad de sufrir un accidente durante manejo se incrementa. Por lo tanto, hasta la fecha han sido propuestos varios esquemas que detectan la emoción del conductor basado en su expresión facial. La mayoría de ellos usan solo un fotograma (una imagen) y operan en condiciones restringidas que rara vez se presentan en condiciones reales de manejo. Con el fin de poder resolver este problema, este artículo presenta un algoritmo para el reconocimiento de la emoción del conductor basado en sus expresiones faciales, en el cual a partir de la secuencia de cuadros de video se extraen vectores de características temporales usando los puntos relevantes del rostro. Los vectores de características extraídos se introducen en diferentes clasificadores, tales como Máquina de Vector-Soporte (SVM) y K-vecinos más cercanos (KNN) para la comparación de su funcionamiento.

Palabras Clave: Emoción de conductor, Expresión facial, Puntos relevantes de rostro, características temporales, SVM, KNN

Abstract

An important factor that influences car accidents is driving under non-optimal conditions, such as stress, anger, fear, depression, among others, in which the drivers face up the situations that increase the probability of an accident while driving. Therefore, various the driver's emotion detector based on the facial expression have been proposed to date. However, most of them use only one video frame (an image) and operate in restricted conditions that are rarely encountered in real driving. To be able to solve this problem, this paper presents an algorithm for driver's emotion recognition using the facial expression, in which vectors of temporal characteristics are extracted from the video sequence using the Face Landmarks of each frame. The extracted feature vectors are fed into different classifiers, such as the Support Vector Machine (SVM) and the K-Nearest Neighbors (KNN) for performance comparison.

Keywords: Driver's emotion, Facial expression, Face Landmarks, Temporal features, SVM, KNN

1. Introducción

Uno de los factores importantes que influyen en los accidentes automovilísticos es el manejar bajo condiciones mentales y emocionales no óptimas, tales como estrés, ira, miedo, depresión entre otros, en las cuales los conductores se enfrentan a situaciones negativas que incrementan la probabilidad de sufrir un accidente durante el manejo. En general, los conductores con estrés están distraídos y como resultado pierden el correcto criterio para analizar su alrededor (Hina et al. 2017), lo cual podría ocasionar un accidente, tomando actos erróneos. En (Fujii 2014), se mostró una relación clara entre la velocidad de manejo y las emociones de conductor. Los conductores con ira o tristeza suelen incrementar inconscientemente la velocidad de automóvil. Este fenómeno no ocurre bajo otras emociones, tales como felicidad, miedo y neutral (Hongyu et al. 2019, Yoshimoto et al. 2021). Considerando la existencia de una relación clara entre el estado emocional del conductor y el factor de riesgo en la conducción del automóvil, se han propuesto varios trabajos.

*Autor para la correspondencia: H Perez-Meana
Correo electrónico: hmperezm@ipn.mx (H. Perez-Meana), inge.marcos.sr@gmail.com (M. Sánchez-Ruiz), jonathan123987j@gmail.com (J. Flores-Monroy), eescamillah@ipn.mx (E. Escamilla-Hernandez), mnakano@ipn.mx (mnakano@ipn.mx)

Historial del manuscrito: recibido el 02/05/2022, última versión-revisada recibida el 10/08/2022, aceptado el 24/08/2022, publicado el 05/10/2022
DOI: https://doi.org/10.29057/icbi.v10iEspecial4.9142

Dependiendo de las señales de entrada para determinar la emoción de los conductores, los métodos se pueden clasificar en dos categorías: El método basado en señales fisiológicas, tales como la señal de un Electrocardiograma (ECG) y la señal de un Electroencefalograma (EEG) (Yoshimoto et al. 2021); y el método basado en imágenes del rostro de los conductores (Zadobrischi et al. 2020), (Malaescu, et al. 2019). Las señales fisiológicas son obtenidas a través de sensores puestos en el conductor, tales como un casco para EEG y un reloj de pulsera para EEG (Yoshimoto et al. 2021). Aunque la exactitud de la detección de la emoción es superior que el método basado en las imágenes de rostro, este método requiere forzosamente sensores durante el manejo para determinar la emoción del conductor.

Mientras que el método basado en la expresión facial del rostro del conductor requiere una cámara web o la cámara que trae un teléfono celular para captar imagen de rostro del conductor. El método propuesto por (Zadobrischi et al. 2020) usa Redes Neuronales Convolucionales basado en región (R-CNN) para segmentar la región de ambos ojos y la región de la boca, y posteriormente cada región se introduce a un arreglo de Redes Neuronales Convolucionales para su clasificación entre seis emociones, tales como: triste, ira, sensible, feliz, excitado y asustado. Según (Zadobrischi et al. 2020), la mejor precisión obtenida es 78% y la peor precisión es 14%. Los autores de (Malaescu, et al. 2019) presentaron un método basado en CycleGAN para crear artificialmente las imágenes de rostro tomadas por cámaras infrarrojas a partir de las imágenes tomadas por cámaras de luz visible. Usando todas las imágenes, originales y las creadas artificialmente, se entrena una CNN, tipo VGG16, para clasificar imágenes de rostro en siete emociones. La razón de crear imágenes de luz infrarroja es la falta de imágenes naturales en el dominio infrarrojo. La exactitud de la clasificación es 74.58%, la cual es 10% más alta que el sistema usando solamente imágenes originales sin usar imágenes artificiales.

Se considera que la expresión facial de una persona no se puede determinar fácilmente desde una sola imagen, ya que los mismos seres humanos entendemos la emoción de una persona, observándola durante un lapso, aunque la duración sea menor de un segundo. Considerando esta observación, en este artículo se propone un método de extracción de características temporales usando los puntos relevantes del rostro en una secuencia de video. Como puntos relevantes, se usó *MediaPipe Mesh* (Kartynnik et al. 2019), la cual entrega una maya con 468 puntos relevantes del rostro. Los puntos tienen las coordenadas de 3D, los cuales ayudan a distinguir más fácilmente las diferentes emociones. Para que las características sean invariantes ante variaciones reales que ocurren durante el manejo, las características son normalizadas. Las variaciones consideradas son la distancia entre la cámara y el rostro del conductor, la condición de luminosidad, orientación del rostro (el conductor está volteado izquierdo o derecho). Como expresiones faciales, se consideran las siguientes tres emociones: enojado, feliz y sorprendido.

Los vectores de características a lo largo del tiempo se introducen a diferentes clasificadores, dentro de los que se tienen: el SVM, KNN, clasificadores basados en árboles, clasificadores ensamblados, entre otros. Después del entrenamiento de los clasificadores, se obtienen las exactitudes proporcionadas por diferentes clasificadores. El SVM y KNN ponderado proporcionaron mejores resultados de clasificación, siendo 88.9% y 89.9%, respectivamente.

El resto de este artículo está organizado en la siguiente forma: En la sección 2, se describe el sistema propuesto que consiste en extracción de características, clasificador y toma de decisiones. En la sección 3, se proporcionan los resultados obtenidos, y finalmente se concluye el trabajo junto con los trabajos futuros en la Sección 4.

2. Sistema propuesto

El sistema propuesto para el reconocimiento del estado de ánimo del conductor se muestra mediante un diagrama a bloques en la Figura 1. Como se puede observar de la figura, a cada trama de la secuencia de video, se aplica la localización de rostro usando *MediaPipe Face* (Bazarevsky et al. 2019) y su normalización. De cada rostro localizado y delimitado se extraen los puntos relevantes usando *MediaPipe Mesh* (Kartynnik et al. 2019). Usando los puntos de labios y ojos se calculan cuatro distancias con las que se forman los vectores de características, los cuales se pueden usar para la identificación del estado de ánimo de los conductores. Finalmente, los vectores de características se introducen a un clasificador.

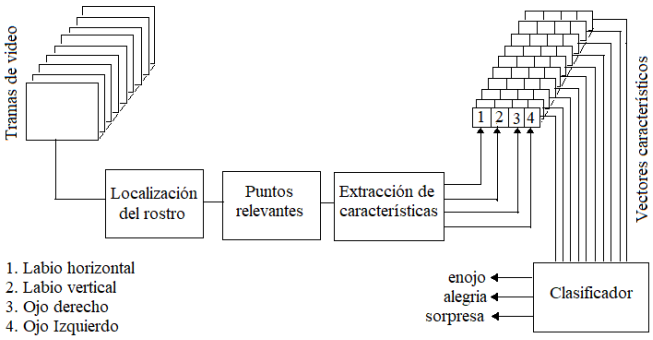


Figura 1: Diagrama bloque del sistema propuesto

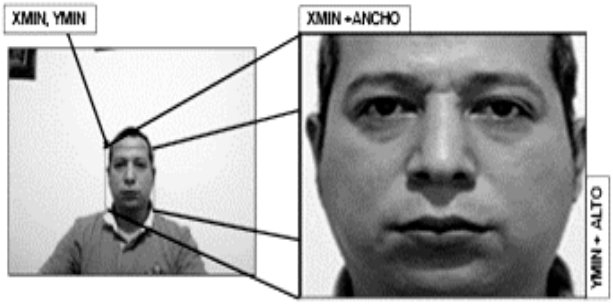


Figura 2: El proceso de normalización de región de rostro.

2.1 Extracción de características

Como se mencionó anteriormente, a cada trama de video se le aplica el *MediaPipe Face* (Bazarevsky et al. 2019), para detectar la región de rostro usando "*bounding-box*", el cual ofrece un punto de referencia en la parte superior derecha en un plano en 2D del rostro la cual se denotará como (xmin, ymin), posteriormente se delimitará el ancho y alto del rostro la cual será la nueva imagen para analizar. Este proceso se muestra en la figura 2. La parte izquierda de la figura muestra el rostro detectado usando "*bounding-box*" de una trama de video y la parte derecha es la imagen del rostro delimitada.

Una vez que la región de rostro fue detectada, delimitada y normalizada, el *MediaPipe Mesh* (Kartynnik et al. 2019) se aplica a la región. Como se había mencionado anteriormente, *MediaPipe Mesh* detecta los 468 puntos relevantes del rostro en el espacio 3D, como se muestra en la figura 3, en la cual los 468 puntos se localizan en los vértices que forman los triángulos de la maya. La aplicación de *MediaPipe Mesh* a la imagen del rostro detectado se muestra en la figura 4.

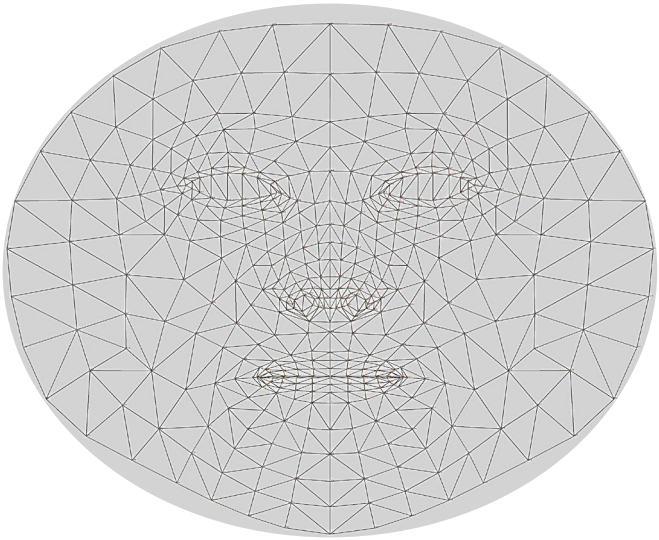


Figura 3: Maya facial con 468 puntos relevantes.

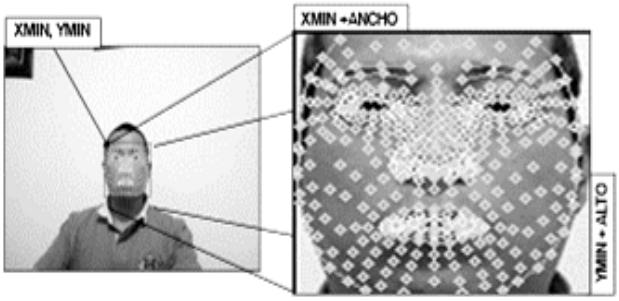


Figura 4: Maya facial aplicada a Fig.3.

Como características que permiten distinguir los estados de emoción de conductores, se consideraron y analizaron cuatro medidas, las cuales son: (1) la distancia vertical entre labios superior e inferior, (2) la distancia horizontal entre dos puntos extremos de los labios, (3) la distancia entre parpados arriba y abajo en ojo izquierdo y (4) la distancia entre parpados arriba y abajo del ojo derecho. La figura 5 muestra las distancias horizontal y vertical de labios, mientras la figura 6 muestra las distancias de los ojos izquierdo y derecho. Todas las distancias han sido calculadas usando la distancia euclidiana en 2D.

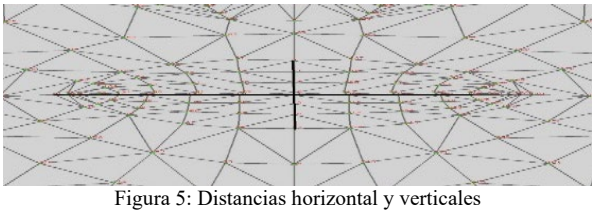
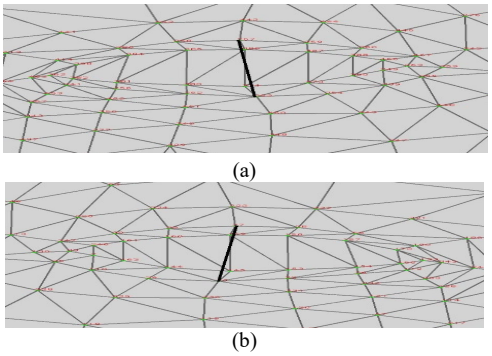


Figura 5: Distancias horizontal y verticales



(a)

(b)

Figura 6: Distancias entre parpados superior e inferior. (a) ojo izquierdo y (b) ojo derecho.

La figura 7 muestra la obtención de cuatro características de un rostro.

Figura 7: Obtención de cuatro características de un rostro.

2.2 Clasificador

Como clasificador del sistema propuesto, se puede usar cualquier clasificador o método de aprendizaje automático, ya que el número de elementos de cada vector de características es pequeño. Se probaron los siguientes clasificadores: SVM con diferentes "*kernels*", tales como "*kernel*" lineal, polinomial de segundo y tercer orden, función Gaussiana con diferentes valores de varianza, KNN con diferente número de vecinos K, clasificadores basados en árboles y clasificadores combinados. Dentro de todos clasificadores probados, los clasificadores SVM y KNN mostraron mejores rendimientos. Por lo tanto, en esta sección, se explica brevemente estos dos clasificadores.

- La Máquina de Vector-Soporte (SVM): El SVM es un clasificador poderoso, el cual está basado en el concepto de margen (Cortes y Vapnik, 1995). La idea básica del SVM es obtener una frontera de separación de clases donde la distancia (margen) entre la frontera y el elemento más cercano a la frontera sea máxima. El elemento de cada clase que se encuentra más cerca de la frontera se llama el vector de soporte. Los cuatro "*kernels*" usados están dados por (1) y (2).

$$\boldsymbol{ker}(\boldsymbol{x_i}, \boldsymbol{x_j}) = \left(\boldsymbol{x_i^T x_j} + \boldsymbol{C}\right)^{\boldsymbol{d}} \quad (1)$$

donde d es el orden, cuando *d*=1, este "*kernel*" presenta "*kernel*" lineal, *d*=2 presenta "*kernel*" polinomial de segundo orden (kernel cuadrático) y *d*=3 presenta "*kernel*" polinomial de tercer orden (*kernel* cúbico). El "*kernel*" Gaussiano está dado por (2).

$$ker(x_i, x_j) = exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \quad (2)$$

donde σ es desviación estándar de la función Gaussiana.

Este clasificador fue desarrollado originalmente para la clasificación binaria, por lo tanto, cuando el número de clases incrementa, su rendimiento disminuye.

- K-vecinos Más Cercanos (KNN): El KNN está basado en el concepto de similitud entre un dato de entrada con los datos etiquetados previamente. Se asigna una clase al dato de entrada tomando cuenta los *K* elementos con mayores similitudes, aplicando decisión por mayoría. Aunque el concepto de este clasificador es muy simple, en general ofrece muy buen desempeño para cualquier tipo de clasificación. Los hiper-parámetros importantes de este clasificador son el número de vecinos *K* y la métrica de distancia que se usa para medir la similitud. En el sistema propuesto, se usa la distancia euclidiana. Como una variante de KNN, existe un clasificador KNN ponderado, en el cual las distancias entre la entrada y cada uno de *K* vecinos son considerados para la toma de decisión, como se muestra (3).

$$\tilde{c} = arg\max_{c}\left[\sum_{i=1}^{K} w_{n_i} I_c(n_i)\right] \quad (3)$$

donde $\tilde{c}$ es la clase asignada al dato de entrada, w_{n_i} es peso de i-esimo vecino n_i, i=1,..,*K*, *K* es el número de vecinos y $I_c(n_i)$ es una función dada por (4).

$$I_c(z) = \begin{cases} 1 & si\ z\ pertenece\ a\ clase\ c \\ 0 & en\ otro\ caso \end{cases} \quad (4)$$

3. Resultados experimentales

En esta sección, se describe la base de datos usada para obtener los resultados, posteriormente se muestran algunos ejemplos de extracción de características y finalmente los resultados de clasificación de emociones usando los dos clasificadores con mejor funcionamiento: SVM y KNN.

3.1 Base de datos usada

Para entrenar los clasificadores, se usó una base de datos pública llamada "*Indian Semi Acted Facial Expression Database*" ISAFE (Singh y Benedict 2019). Esta base de datos contiene secuencias de video de 44 voluntarios entre 17 y 22 años. La expresión facial de cada voluntario cuando él o ella está viendo diferentes escenas que provocan ocho diferentes tipos de emociones, incluyendo estado neutro. Las siete emociones son: feliz, triste, sorprendido, disgusto, miedo, enojo e incertidumbre. Para la construcción de nuestra base de datos, se seleccionaron tres emociones, las cuales son feliz, sorprendido y enojado. Usando el método de la extracción de características propuesto, se obtienen las características temporales de las tres emociones. El número de datos de las características para las tres emociones es de 340 por cada emoción.

3.2 Ejemplos de vectores de características

En esta Sección, se muestran algunos ejemplos de los vectores de características extraídos cuando el rostro presenta diferentes expresiones faciales. La figura 8 muestra un rostro con el estado de emoción "Feliz", y la figura 9 muestra cuatro características a lo largo del tiempo durante el estado de emoción "Feliz".

Figura 8: Un ejemplo de rostro con el estado de emoción "Feliz".

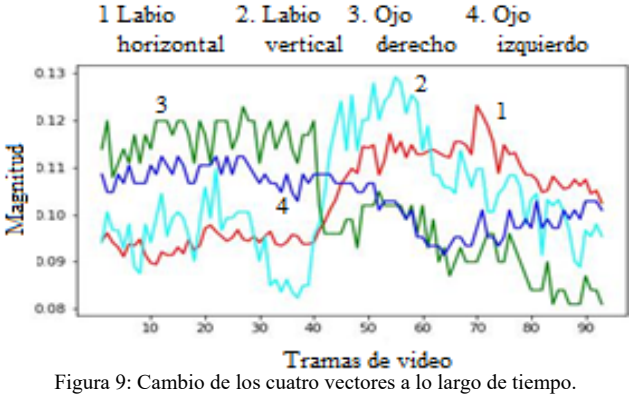


Figura 9: Cambio de los cuatro vectores a lo largo de tiempo.

Las figuras 10 y 11 muestran un ejemplo del estado de emoción "sorprendido", siendo la figura 10 una trama de video y figura 11 el cambio de las cuatro características extraídas del rostro.

Figura 10: Un ejemplo de rostro con el estado de emoción "Sorprendido".

Así mismo las figuras 12 y 13 muestran un ejemplo del estado ánimo "enojado" y las cuatro características extraídas de las tramas de video correspondientes.

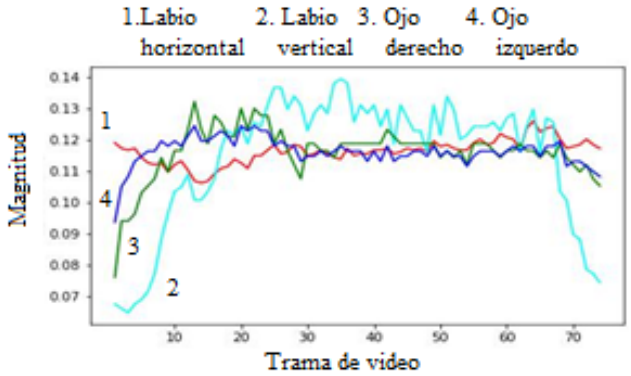


Figura 11: Cambio de los cuatro vectores a lo largo de tiempo.


Figura 12: imagen de rostro con el estado de "enojado"

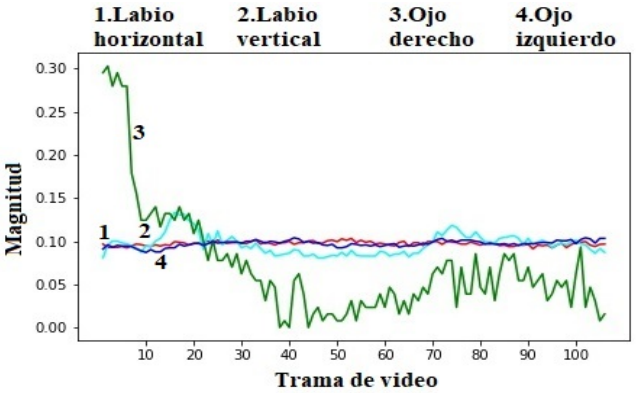


Figura 13: Las 4 características cuando el estado es "enojado"

Como se puede observar, los cuatros vectores tienen cambios significativos durante los tres estados "feliz, "enojado", "sorprendido", dependiendo de la emoción se obtienen diferentes medidas las cuales serán implementadas en los clasificadores posteriormente en la aplicación del clasificador.

3.3 Resultados de reconocimiento

Después de evaluar diversos clasificadores con diferentes hiper-parámetros, se seleccionaron los dos clasificadores que proporcionaron mejor funcionamiento en la de detección de emociones, los cuales son el SVM con "*kernel*" Gaussiano con desviación estándar σ=0.5 y el KNN ponderado con el número de vecinos *K*=10.

La figura 14 muestra la matriz de confusión de clasificación realizada por el SVM. Como se puede observar de la figura, la emoción "feliz" se confunde más con otras emociones, esto debido a que la expresión facial de felicidad varía de persona a persona. La figura 15 muestra la distribución de dos elementos del vector de características, Labio-horizontal y Labio-vertical clasificadas para los tres estados de emoción, "feliz", "enojado" y "sorprendido", con "**o**" para las clasificaciones acertadas y en "**x**" para las incorrectas, teniendo los siguientes colores para cada estado, rojo para el estado "feliz", anaranjado para el estado "sorprendido" y azul para el estado "enojado". En la misma forma, las figuras 16 y 17 muestran la matriz de confusión y la distribución de las dos características antes mencionadas que corresponden a los resultados del KNN.

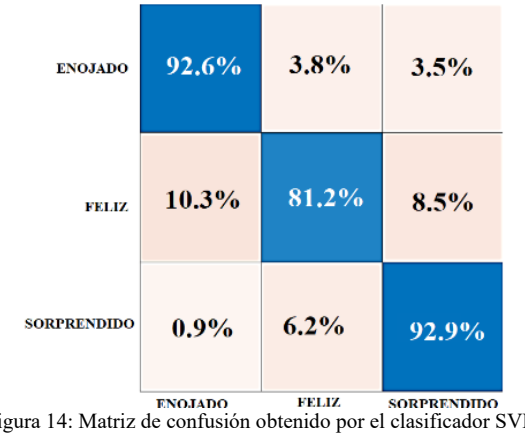


Figura 14: Matriz de confusión obtenido por el clasificador SVM

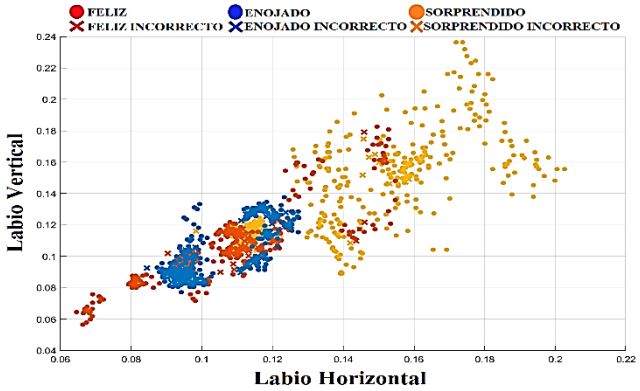


Figura 15: Distribución de dos elementos en las tres clases, con clasificaciones correctas e incorrectas

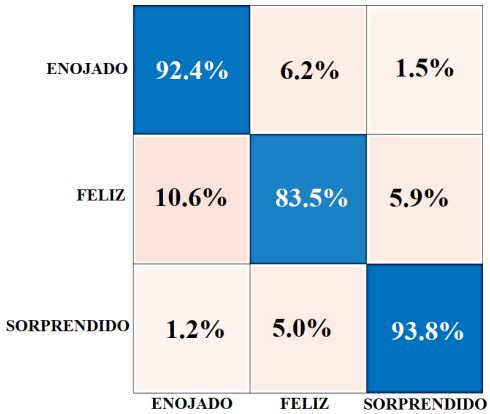


Figura 16: Matriz de confusión obtenido por KNN ponderado.

En la Tabla 1, se puede observar la comparativa de los clasificadores que proporcionaron mejores resultados y los parámetros de desempeño de ambos modelos, teniendo como resultado una pequeña diferencia en *Accuracy* en los modelos SVM y kNN del 1%. Así mismo se puede observar en la gráfica de la figura 19 la comparativa entre algunos de los

modelos de clasificadores usados, con lo cual se puede observar que hay mejores resultados en los dos clasificadores SVM Gaussiano y KNN ponderado.

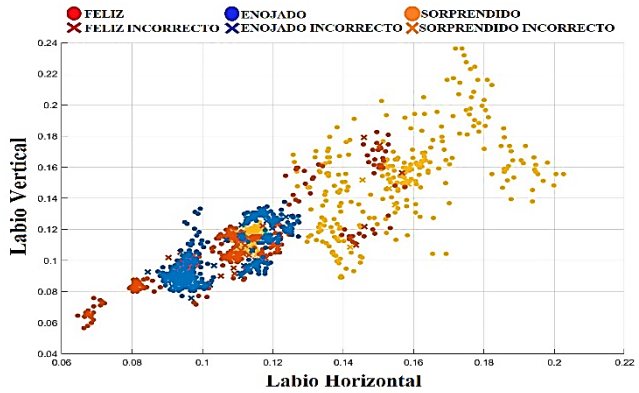


Figura 17: Distribución de dos elementos en las tres clases, con clasificaciones correctas e incorrectas

Tabla 1: Resultados de SVM y KNN

METRICA	SVM σ=0.5			KNN Ponderado k=10		
PRECISION	81.2%	92.7%	93%	93.5%	92.4%	93.8%
RECALL	89%	89.2%	89%	88.2%	88.7%	92.7%
F1-SCORE	85%	92%	91%	85.8%	90.5%	93.3%
SPECIFITY	90.8%	96%	96%	92%	96%	97%
ACCURACY		88.9%			89.9%	

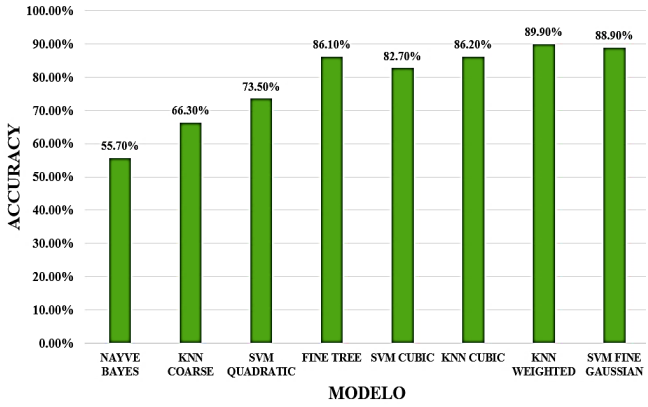


Figura 18: Comparación de exactitud con diferentes clasificadores

4. Conclusiones

En este artículo se propone un esquema para detectar el estado de emoción usando la expresión facial captada por una webcam. El sistema consiste en la detección del rostro usando *Mediapipe Face* (Bazarevsky et al. 2019), el proceso de normalización para que el sistema sea robusto a variaciones de la distancia entre la cámara y el rostro del conductor. Seguidamente al rostro detectado, se le aplica *Mediapipe Mesh* (Kartynnik et al. 2019) para obtener los puntos relevantes, en los cuales se extrajeron cuatro métricas que distinguen tres emociones, "Enojado", "Feliz" y "Sorprendido".

Como clasificadores, se probaron SVM con diferentes "*kernels*", KNN con diferentes números de vecinos *K,* y otros clasificadores, tales como clasificadores basados en árboles, clasificadores ensamblados. Dentro de los clasificadores usados, SVM con "*kernel*" Gaussiano y el KNN ponderado con el número de vecinos *K*=10 mostraron mejores resultados, obteniendo las exactitudes 88.9% y 89.9%, respectivamente. Las cuatro características que se obtuvieron por medio de los puntos relevantes del rostro fueron importantes para obtener esta alta tasa de acierto en la clasificación.

Como trabajos futuros que se puedan derivar de este artículo se tienen los siguientes: (1) Aumentar el número de emociones que se pueden detectar. Actualmente tres emociones son detectadas, sin embargo, otras emociones, tales como miedo o disgusto, tristeza, son importantes para evitar posibles accidentes. (2) Analizar otras métricas usando puntos relevantes de los rostros. En este trabajo, se usaron cuatro métricas relacionadas a labios y ojos, sin embargo, los movimientos de nariz o frente están relacionados con algunas expresiones faciales. (3) La exploración de redes neuronales profundas, sobre todo redes neuronales recurrentes profundas. En este trabajo, se optó por el uso de clasificadores, tales como el SVM, KNN entre otros, ya que el número disponible de datos de entrenamiento es muy limitado. Sin embargo, si se construye una base de datos propia grabando videos, en lugar de depender de las bases de datos públicas, se tendría la posibilidad de obtener un conjunto de entrenamiento con suficiente número de datos para entrenar redes neuronales recurrentes profundas. Se considera que las redes neuronales recurrentes, tales como "*Long Short Term Memory*" (LSTM) pueden ser una alternativa, ya que estas redes utilizan características de secuencia temporal para la clasificación.

Referencias

Bazarevsky, V., Kartynnik, Y., Vakunov, A., Raveendran, K., Grundmann, M. (2019). BlazeFace: Sub-millisecond Neural Face Detection on Mobile GPUs. Proceedings of Computer Vision & Pattern Recognition. arXvis:1907.05047v2

Cortes, C., Vapnik, V. (1995). Support vector networks. Machine Learning. 20, 237-297.

Fujii, Y. (2014). Study on driver's anger emotions and their coping behaviors", The Journal of Clinical Research Center for Child Development and Educational Practices

Hima, MD., Guan, H., Ramdane-Cherif, Amar. (2017). Novel approaches in human-vehicle interaction interface of a vehicle driving assistance system.

Hongyu, H., Zhou, X., Zhu, Z., Wang, Q., Xiao, H. (2019). A Driving simulator study of young driver's behavior under angry emotion, SAE Technical Paper 2019-01-0398, 2019, https://doi.org/10.4271/2019-01-0398.

Karthynnik, Y., Ablavatski, A., Grishchenko, I., Grundmann, M. (2019). Real-time facial Surface geometry from monocular video on mobile GPUs. CVPR Workshop on Computer Vision for Augmented and Virtual Reality. https://doi.org/10.48550/arXiv.1907.06724

Malaescu, A., Duju, L-C., Sultana, A., Filip, D., Ciuc M. (2019). Improving in-car emotion classification by NIR database augmentation. Proceedings - 14th IEEE International Conference on Automatic Face and Gesture Recognition, 8756628.

Singh, S., Benedict, S. (2019). Indian semi-acted facial expression (iSAFE) dataset for human emotions recognition. Advances in Signal Processing and Intelligent Recognition Systems, Communications in Computer and Information Science, 1209.

Yoshimoto, H., Sakai, K., Hiramatsu, Y., Ito A. (2021). Building a sensor network to measure driver's emotions. 9th Int. Symp. on Computing and Networking Workshops. pp.77-80.

Zadobrischi, E., Cosovanu, L-M., Negru, M., Dimian, M. (2020). Detection of emotional states through the facial expressions of drivers embedded in a portable system dedicated to vehicles. 28th Telecommunications forum TELFOR