

Informe de integración SISO-VID

Elaboración:	Equipo técnico
Entidad Ejecutora:	X-TRA PLUS SOLUCIONES DE ENERGÍA S.A.C
Proyecto:	Desarrollo de un Sistema Adaptativo para la Detección de Somnolencia en Conductores de Transporte Interprovincial idóneo para las características únicas de las Carreteras del Perú mediante Sensado Híbrido utilizando Técnicas de Deep Learning.
Periodo:	Marzo 2021
Fecha:	13 de marzo de 2021

1. Objetivo

Implementar el modulo SISO-VID. Este componente de software es el encargado de detectar somnolencia a partir de una secuencia de video utilizando aprendizaje profundo.

2. Introducción

Según *World Health Organization* WHO (2013), 1.24 millones de accidentes de tráfico ocurren cada día. Además, *The National Highway Traffic Safety Administration* (NHTSA), menciona que en USA han ocurrido 153,297 accidentes de tránsito entre el 2011 al 2015, y de estos el 2.4 % fueron causados por conductores con somnolencia. Incluso, 1.25 millones de personas mueren cada año en accidentes de tránsito, 20 a 50 millones han sido heridos o están discapacitados y todo esto ha llegado a costar 518 billones de dólares. Mas alarmante, se predice que los accidentes de tránsito serán la quinta causa mas frecuente de muertes para el 2030 (ASIRT, 2020).

Los principales métodos de detección de somnolencia pueden dividirse en 3 grandes grupos según Ramzan et al. (2019). El primer grupo está conformado por los métodos basados en procesamiento de imágenes y video, estos métodos captan la señal a partir de una cámara y utilizando algoritmos de visión computacional logran detectar la somnolencia. El segundo grupo está conformado por los métodos que utilizan sensores en el auto, estos sensores miden la fuerza de agarre del timón, los cambios en el tiempo de los ángulos de giro e incluso analizan las señales de frecuencia cardíaca. El último grupo, es muy intrusivo y se basan en medidas fisiológicas como el *electroencephalogram* (EEG), *electrooculogram* (EOG), *electromyogram* (EMG) y *electrocardiogram* (ECG).

SISO-VID es un método basado en comportamiento, este módulo toma como entrada una secuencia de video y detecta si alguna escena de la secuencia de video presenta somnolencia. En resumen, se ha utilizado detección de rostros con una red neuronal *Single Shot Detector* (SSD), luego sobre el rostro detectado se utilizó la red neuronal *Inception* para detectar somnolencia. El método propuesto logró un 87 % de accuracy sobre la base de datos SISO-IMG, esta base de datos fue creada específicamente para este proyecto (para mas detalles, puede revisar el informe de base de datos).

3. Metodología de SISO-VID

SISO-VID es un componente de software basado en aprendizaje profundo para la detección de somnolencia tomando como entrada una secuencia de video. En la Figura 1, presentamos la metodología

utilizada de SISO-VID. El método propuesto toma una secuencia de video como entrada, luego extrae los fotogramas, por cada fotograma se detecta el rostro utilizando la red neuronal SSD, para finalmente utilizar la red Inception sobre el rostro detectado para determinar si exista somnolencia.

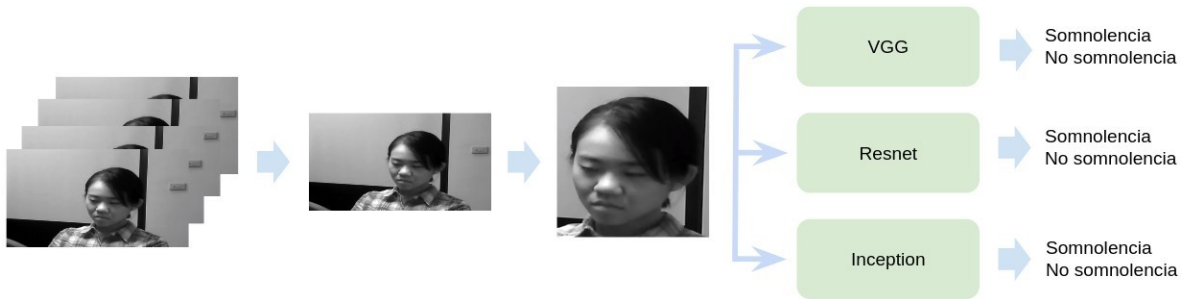


Figura 1: Metodología utilizada por SISO-VID.

3.1. Muestreo de fotogramas

El primer paso consiste en hacer un muestreo de la secuencia de video, es decir, tomando como entrada una secuencia de video, se extraen solo algunos fotogramas que sirvan de entrada a los pasos siguientes (ver Figura 2). Se decide hacer este muestreo porque procesar todos los fotogramas demandaba mucho tiempo de procesamiento, considerando que SISO-VID opera en una Jetson Nano. Para el proyecto, se optó por hacer un muestreo de tres fotogramas cada segundo.

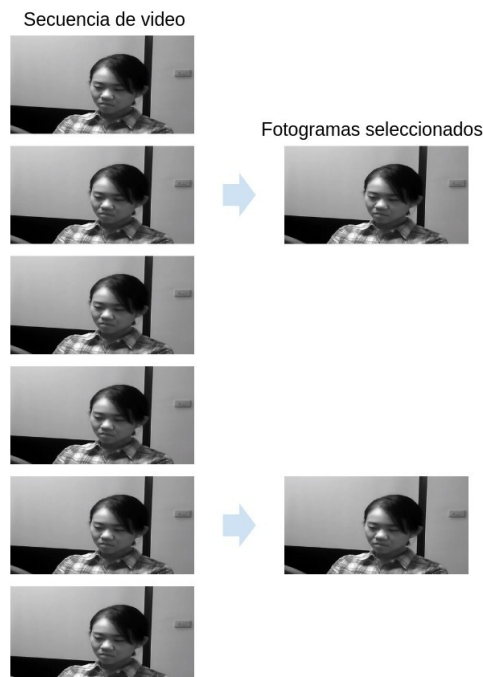


Figura 2: Muestreo de los fotogramas en SISO-VID.

3.2. Detección de rostros

La detección de rostros consiste en obtener la posición de un rostro dada como entrada una imagen. Por ejemplo, en la Figura 3, tenemos una imagen, luego podemos aplicar algoritmos para obtener las coordenadas dentro de la imagen donde existan rostros, mayormente estas coordenadas son representadas con un rectángulo enfocando el rostro.

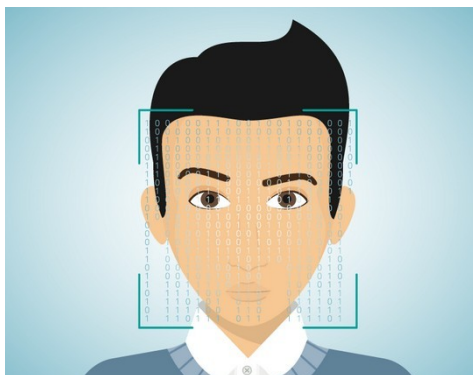


Figura 3: Ejemplo de detección de rostros.

Existen varios algoritmos de detección de rostros, uno de los primeros fue presentado por Viola and Jones (2004). También existen algunas librerías, una de las más utilizadas es *dlib*¹. Recientemente, se está utilizando aprendizaje profundo para la detección de rostros, entre estas tenemos: *Multi Task Convolutional Neural Netowrk* (MTCNN) (Zhang et al., 2016) y una red neuronal ya entrenada en OpenCV 3.0, basada en SSD e implementada en Caffe. Nosotros escogimos la segunda red neuronal, al ser la de mejor acierto.

3.3. Detección de somnolencia

Con cada rostro detectado en la etapa anterior, se evalúa si este presenta o no somnolencia. En esta etapa se utilizó tres redes neuronales distintas: VGG, Inception y Resnet (ver Figura 4). Se escogieron estas redes al tener un buen desempeño en el estado del arte en clasificación de objetos. De las redes neuronales propuestas, el modelo Inception obtuvo los mejores resultados.

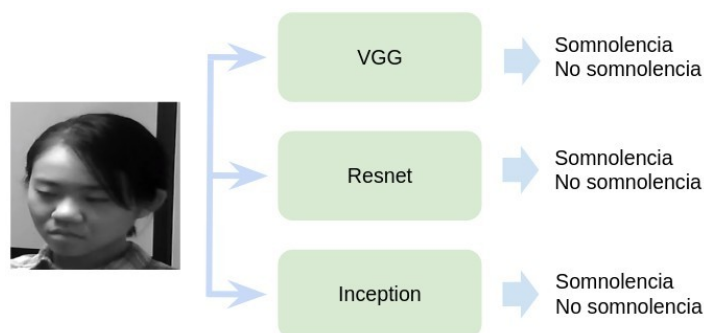


Figura 4: Detección de somnolencia a partir de un rostro utilizando VGG, Resnet e Inception.

La red neuronal Inception Szegedy et al. (2016) en su tercera versión, es un modelo de reconocimiento de imágenes muy utilizado, está logró una exactitud de 78.1 % en la base de datos ImageNet.

¹Librería de *machine learning* y procesamiento de imágenes, desarrollada en C++. Enlace.

Debido a esto, se utilizo esta red neuronal para detectar somnolencia.

La red neuronal Inception-v3 hace uso de unos módulos llamados Inception. Estos actúan como filtros aplicados a un mismo valor de entrada mediante capas convolucionales y de *pooling*. Esto permite sacar provecho de la extracción de patrones que brindan diferentes tamaños en los filtros. Luego, el resultado de estos filtros es concatenado y utilizado como el valor de salida del módulo. Este modelo aumenta el número de parámetros entrenables y la computación requerida, pero mejora considerablemente la exactitud. En la Figura 5, mostramos la arquitectura de esta red.

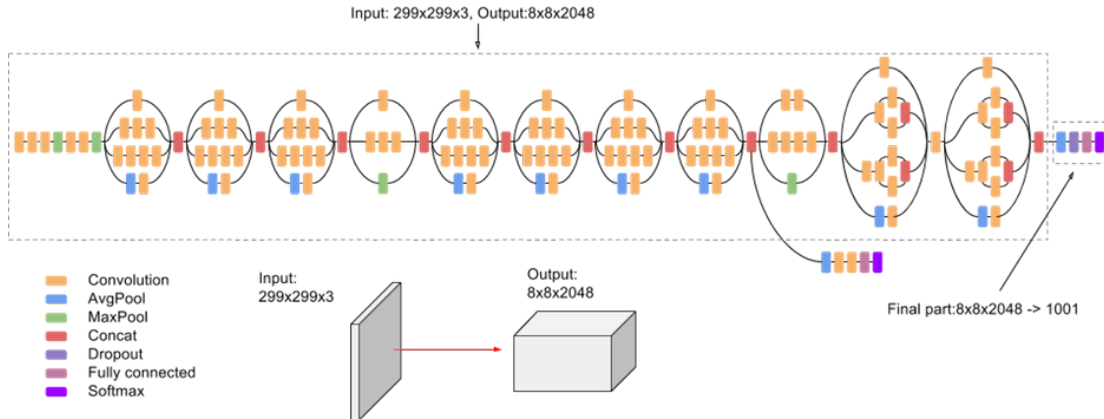


Figura 5: Detección de somnolencia a partir de un rostro utilizando VGG, Resnet e Inception.

3.4. Conteo de parpadeos y bostezos

De manera adicional al método propuesto para la detección de somnolencia utilizando la red neuronal Inception, se ha agregado un método que cuenta la cantidad de parpadeos y bostezos. En esta etapa, se utilizó *dlib* para la detección de los *landmarks* del rostro del conductor, luego con los *landmarks* de los ojos y boca se puede determinar si los ojos están cerrados o si la boca es abierta.

En el caso de la detección de parpadeos, el *Eye Aspect Ratio* (EAR) es calculado con la Ecuación 1 (los landmarks están en la Figura 6), este valor está entre 0(cerrado) y 1(abierto), en nuestro caso, definimos que si este EAR era menor a 0.3, se consideraba como un parpadeo.

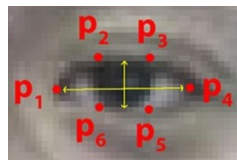


Figura 6: Landmarks de los ojos utilizado para calcular el EAR.

$$EAR = \frac{|p_2 - p_6| + |p_3 - p_5|}{|p_1 - p_4|} \quad (1)$$

En el caso de la detección de bostezos, el *Mouth Aspect Ratio* (MAR) es calculado con la Ecuación 2 (los landmarks están en la Figura 7), en nuestro caso, definimos que si este MAR era mayor a 0.7, se consideraba como un bostezo.

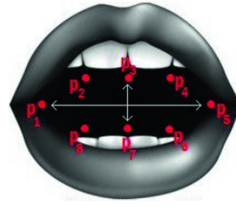


Figura 7: *Landmarks* de los ojos utilizado para calcular el EAR.

$$MAR = \frac{|p_2 - p_8| + |p_3 - p_7| + |p_4 - p_6|}{|p_1 - p_5|} \quad (2)$$

4. Resultados

En la Figura 8, mostramos como es el funcionamiento de SISO-VID evaluado en la base de datos creada para este proyecto. Esta base de datos tiene muestras de conductores reales en su rutina normal de trabajo de una empresa de transporte.



Figura 8: SISO-VID en ejecución en la base de datos creada para el proyecto.

En la Figura 9, mostramos como es el funcionamiento de SISO-VID, pero esta vez evaluado en la base de datos NTHU. La base de datos NTHU, tiene muestras de videos de personas con somnolencia

en una oficina con buenas condiciones de iluminación y calidad de imagen.



Figura 9: SISO-VID en ejecución en la base de datos NTHU.

Finalmente en la Figura 10, mostramos el conteo de parpadeos y bostesos. Si bien esta información no es determinante para la detección de somnolencia, porque la frecuencia de parpadeos no siempre determina si alguien está somnoliento, se utilizará esta información para trabajos futuros.

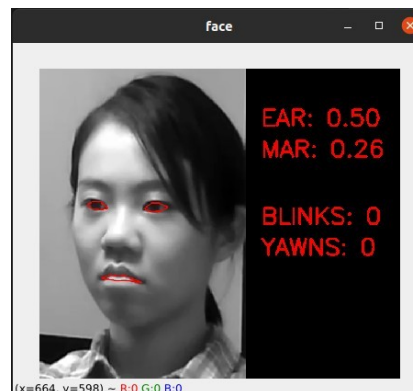






Figura 10: Conteo de parpadeos y bostesos con SISO-VID.

5. Conclusiones

En este informe se ha detallado los algoritmos y métodos utilizados para la implementación de SISO-VID, el cual es el módulo encargado de la detección de somnolencia a partir de secuencias de video utilizando aprendizaje profundo.

El método propuesto se basa en hacer un muestreo preliminar de algunos fotogramas, para reducir el tiempo de procesamiento, luego se aplica detección de rostros, para finalmente hacer la detección de somnolencia con la red neuronal convolucional Inception.

El método propuesto logró una exactitud de 87 % evaluado en la unión de la base de datos NTHU y una creada especialmente para este proyecto. Adicionalmente, se evaluarón otros modelos, pero obtuvieron un menor desempeño (el informe de pruebas detalla estos experimentos).

	Nombre	Cargo	Firma	Fecha
Elaborado por:	Jason Paul Cahua-na Nina	Equipo técnico - Tesis-ta La Salle		12/03/2021
	Karla Mariel Fernández Fabián	Investigador y desarrollador - Área de innovación y desarrollo tecnológico		12/03/2021
	Vicente Enrique Machaca Arceda	Investigador y desarrollador - Área de innovación y desarrollo tecnológico		12/03/2021
Revisado por:	Medardo Delgado Paredes	Investigador La Salle		12/03/2021
	Antonio Simon Bolívar Paredes	Coordinador administrativo del proyecto		12/03/2021
Aprobado por:	Elvis Diego Supo Colquehuanca	Coordinador general del proyecto		12/03/2021

Referencias

- ASIRT (2020). Annual global road crash statistics. <https://www.asirt.org/safe-travel/road-safety-facts/>.
- Ramzan, M., Khan, H. U., Awan, S. M., Ismail, A., Ilyas, M., and Mahmood, A. (2019). A survey on state-of-the-art drowsiness detection techniques. *IEEE Access*, 7:61904–61919.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.
- Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *International journal of computer vision*, 57(2):137–154.
- WHO (2013). World health organization - infographics on global road safety 2013. https://www.who.int/violence_injury_prevention/road_safety_status/2013/facts/en/.
- Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503.