CEA Comité Español de Automática
XLVI JORNADAS AUTOMÁTICA CARTAGENA 2025
UNIVERSIDADE DA CORUÑA

Jornadas de Automática

Fusión de imágenes y señales fisiológicas para modelar al conductor

Fernández, Raúl[a,*], Martín, David[a], de la Escalera, Arturo[a]

[a]*Laboratorio de Sistemas Inteligentes (LSI), Escuela Politécnica Superior, Avda. de la Universidad, 30. 28911 Leganés (Madrid) España*

To cite this article: Fernández, Raúl, Martín, David, de la Escalera, Arturo. 2025. Fusion of images and physiological signals for driver modeling. Jornadas de Automática, 46. https://doi.org/10.17979/ja-cea.2025.46.12121

Resumen

En los niveles intermedios de automatización, los vehículos aún dependen del conductor como respaldo, lo que requiere estimar su disponibilidad cognitiva y física para retomar el control cuando sea necesario. Mientras el uso de imágenes o señales fisiológicas por separado ha demostrado su utilidad, su combinación presenta retos técnicos debido a la disparidad de dimensionalidad. Este trabajo propone una arquitectura de fusión que transforma señales fisiológicas en imágenes para integrarlas con datos visuales mediante autocodificadores entrenados con funciones de pérdida perceptual. La arquitectura se valida inicialmente con señales fisiológicas transformadas en imágenes, analizando distintas técnicas de conversión. De forma separada, se evalúan funciones de pérdida perceptual en autocodificadores aplicados a imágenes, destacando su utilidad para conservar la estructura visual en tareas de reconstrucción. Estos hallazgos refuerzan la viabilidad del enfoque y abren la puerta a futuras ampliaciones que integren datos visuales y evalúen el sistema en contextos más complejos.

Palabras clave: Interacción humano-vehículo, Control compartido, cooperación y nivel de automatización, Automatización y diseño centrado en el ser humano, Modelado fisiológico, Fusión de información y sensores, Redes neuronales, Vehículos autónomos.

Fusion of images and physiological signals for driver modeling

Abstract

At intermediate levels of automation, vehicles still rely on the driver as a backup, which requires estimating their cognitive and physical readiness to resume control when necessary. While the use of separate images or physiological signals has proven useful, combining them presents technical challenges due to the disparity in dimensionality. This work proposes a fusion architecture that transforms physiological signals into images for integration with visual data using autoencoders trained with perceptual loss functions. The architecture is initially validated with physiological signals transformed into images by analyzing different conversion techniques. Additionally, the efficacy of perceptual loss functions in autoencoders applied to images is evaluated, emphasizing their utility in preserving visual structure during reconstruction tasks. These findings underscore the approach's viability and pave the way for future extensions that will integrate visual data and assess the system in more complex scenarios.

Keywords: Human and vehicle interaction, Shared control, cooperation and level of automation, Human-centered automation and design, Physiological Model, Information and sensor fusion, Neural networks, Autonomous Vehicles.

1. Introducción

La progresiva automatización del vehículo está transformando el papel del conductor. Según la clasificación de la Sociedad de Ingenieros Automotrices (SAE), el nivel 0 corresponde a vehículos completamente manuales, mientras que el nivel 1 introduce sistemas de asistencia a la conducción (ADAS). En el nivel 2, el vehículo puede controlar si-

*Autor para correspondencia: raulfern@ing.uc3m.es

multáneamente la dirección y la velocidad, aunque el conductor debe permanecer atento para intervenir en cualquier momento. A partir del nivel 3, el vehículo es capaz por sí solo de realizar la mayoría de las tareas de conducción, aunque todavía requiere intervención humana en determinadas situaciones. En el nivel 4, el vehículo es completamente autónomo, pero solo en un entorno de operación específico; finalmente, el nivel 5 representa la autonomía total, sin necesidad de intervención humana bajo ninguna circunstancia (SAE International, 2021).

En los niveles 2 y 3 de automatización, el sistema puede gestionar ciertas funciones de conducción, pero aún se requiere la intervención del humano para resolver situaciones que exceden sus capacidades. Esto convierte el rol del conductor en intermitente y hace que la colaboración humano-vehículo sea más necesaria que nunca (Marcano et al., 2020). Este control compartido exige que el conductor esté en condiciones de retomar el control en un tiempo razonable, dependiendo del contexto y del nivel de exigencia de la situación. A diferencia de la conducción manual, en estos escenarios el usuario puede desconectarse de la tarea principal si las condiciones lo permiten y es seguro hacerlo (Weigl et al., 2023). Esta desconexión introduce incertidumbre sobre su disponibilidad cognitiva y física ante una petición inesperada de control. Por ello, surgen nuevas preguntas clave: ¿está el humano preparado para retomar el control?, ¿cuánto tiempo necesita para reorientar su atención?, ¿cómo debe comunicarse el vehículo para garantizar una transición fluida? Estas cuestiones, aún sin resolver, son clave para lograr una interacción segura entre el vehículo y la persona.

Para responder a estas preguntas, es necesario contar con herramientas que permitan evaluar el estado del conductor de forma precisa, continua y lo menos intrusiva posible. Una forma de analizar el comportamiento del conductor es mediante el uso de imágenes y técnicas de visión por computador (Kazemi et al., 2025). Otra es emplear señales fisiológicas, como la actividad electrodérmica de la piel, la frecuencia cardíaca o la temperatura corporal, que reflejan el estado del sistema nervioso autónomo y proporcionan información útil sobre el nivel de atención, la fatiga o el estrés del conductor (Deng et al., 2024). Ambas fuentes ofrecen ventajas complementarias, y su integración en una arquitectura de fusión multimodal se presenta como una estrategia prometedora (Wang et al., 2024). Muchas tecnologías de adquisición fisiológica son intrusivas y afectan a la conducción, pero dispositivos como las pulseras inteligentes permiten recoger datos de forma cómoda. Aunque presentan limitaciones, ofrecen una opción viable para integrar señales fisiológicas sin comprometer la experiencia del usuario.

Una de las principales dificultades en la implementación de una arquitectura de fusión multimodal es la diferencia en la dimensionalidad de los datos. Mientras que una señal fisiológica muestreada a 30 Hz proporciona decenas de datos por segundo, un vídeo en alta definición (1080×970 píxeles, 30 fps) puede generar más de 90 millones de datos por segundo. Esta disparidad plantea retos importantes para los modelos de integración. En este trabajo se propone una solución basada en autocodificadores, que permiten reducir la dimensionalidad de la información visual, preservando los aspectos esenciales para el modelado conjunto. Además, se emplean técnicas de conversión de señales en imágenes, lo que permite aumentar la dimensionalidad de las señales fisiológicas y hacerlas compatibles con los métodos de procesamiento visual.

2. Autocodificadores

Los autocodificadores o *autoencoders* son una arquitectura de redes neuronales diseñada para aprender a reconstruir su propia entrada. Su funcionamiento se basa en transformar los datos de entrada a través de una serie de capas, donde la información se va comprimiendo y luego expandiendo, obligando a la red a capturar las características más relevantes de los datos.

Aunque el entrenamiento de un autocodificador utiliza técnicas propias del aprendizaje supervisado (como *backpropagation*), no requiere datos etiquetados, por lo que se encuadra dentro del aprendizaje no supervisado.

Un autocodificador puede definirse matemáticamente como el aprendizaje conjunto de dos funciones:

- Una función de codificación, o *Encoder*, $E : \mathbb{R}^n \to \mathbb{R}^d$, que transforma una entrada $\mathbf{x} \in \mathbb{R}^n$ en una representación latente $\mathbf{z} = E(\mathbf{x}) \in \mathbb{R}^d$, realizando un mapeo de $n \to d$.
- Una función de decodificación, o *Decoder*, $D : \mathbb{R}^d \to \mathbb{R}^n$, que intenta reconstruir la entrada original a partir de la representación latente: $\mathbf{x}' = D(\mathbf{z}) = D(E(\mathbf{x}))$, es decir, mapeando de $d \to n$.

El objetivo del entrenamiento es encontrar los parámetros de la función de codificación E y de la función de decodificación D, denotados por θ_E y θ_D respectivamente, que minimicen una función de error $\mathcal{L}$, la cual cuantifica la diferencia entre la entrada original $\mathbf{x}$ y su reconstrucción $\mathbf{x}'$. La forma concreta de la función de pérdida $\mathcal{L}$ depende del tipo de datos y del propósito del modelo, pero siempre tiene como objetivo que la reconstrucción sea lo más fiel posible a la entrada. Formalmente, esto se expresa mediante la ecuación (1).

$$\min_{\theta_E, \theta_D} \mathcal{L}(\mathbf{x}, \mathbf{x}') \tag{1}$$

Los autocodificadores son redes neuronales muy versátiles que se utilizan en una amplia variedad de tareas, como la generación de imágenes, la eliminación de ruido en señales y, en el marco de este trabajo, la reducción de la dimensionalidad (Bank et al., 2023).

Esta capacidad se basa en la estructura de cuello de botella de la red: la entrada, de dimensión n, se comprime a una representación latente de menor dimensión d al pasar por el codificador. Este proceso obliga al modelo a conservar únicamente la información más relevante, de modo que el decodificador sea capaz de reconstruir la entrada original a partir de esta representación comprimida. De este modo, el codificador actúa como un extractor automático de características y una herramienta eficaz para la reducción de la dimensionalidad.

2.1. Pérdida Perceptual

En tareas de reconstrucción de imágenes, es común utilizar funciones de pérdida elemento a elemento (*element-wise loss*), en las que cada unidad de salida intenta reproducir exactamente su correspondiente valor objetivo. En el contexto de

visión por computador, esto se traduce normalmente en pérdidas definidas a nivel de píxel (*pixel-wise loss*), que penalizan las diferencias individuales entre los valores de cada píxel de la imagen original y su reconstrucción. Sin embargo, este tipo de pérdida presenta limitaciones importantes: no captura las relaciones estructurales entre los distintos elementos de la imagen y asigna la misma importancia a todos los píxeles, aunque algunas regiones puedan ser más relevantes para la percepción visual humana (Pihlgren et al., 2020).

Para abordar estas limitaciones, se emplea la pérdida perceptual (*perceptual loss*), que evalúa la calidad de la reconstrucción no a partir de los valores individuales de los píxeles, sino mediante características extraídas por una red neuronal externa conocida como Red de Pérdida Perceptual (*Perceptual Loss Network*). En lugar de minimizar diferencias directas en la imagen, esta técnica calcula la diferencia entre las salidas de la red de pérdida al procesar tanto la imagen original como su reconstrucción. De esta forma, se obtiene una medida de similitud perceptiva, más alineada con cómo los humanos interpretan la información visual. En este trabajo, se aplican diferentes funciones perceptuales (YOLOv8, VGGNet-16, DeepLabv3-ResNet50), cuya comparación visual se presenta en la Sección 5.2. La Figura 1 muestra la estructura tradicional del autocodificador y cómo se utiliza la red de pérdida perceptual f para computar el error $f(x) - f(x')$ y guiar el aprendizaje del autocodificador.

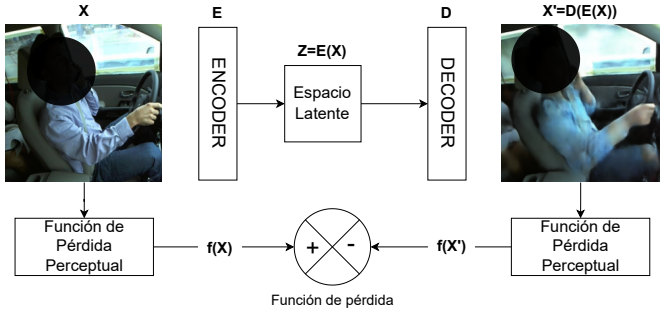

Figura 1: Representación del cálculo de la pérdida perceptual, donde f es la Red de Pérdida Perceptual preentrenada que extrae características de la imagen original (X, del dataset State-Farm Distracted Driver Detection (Montoya et al., 2016)) y la reconstrucción del autocodificador X'. La cara está censurada para proteger la privacidad.

3. Conversión de señal a Imagen

En el contexto de la fusión multimodal para el modelado del estado del conductor, es fundamental que las señales fisiológicas se integren de forma coherente con fuentes visuales como imágenes del rostro o del cuerpo. Dado que estas señales son unidimensionales, cada instante de tiempo refleja un único valor medido, su transformación en imágenes bidimensionales permite alinear su formato con los datos visuales y facilitar una fusión eficiente. Además, esta conversión preserva la dinámica temporal subyacente de las señales, esencial para detectar patrones fisiológicos como picos o transiciones.

3.1. *Preprocesamiento de la señal*

Las imágenes generadas son cuadradas, con ancho y alto iguales y un solo canal. Su dimensión está determinada por el producto de la frecuencia de muestreo de la señal y la longitud de la ventana de trabajo. La configuración de diversos parámetros, ajustados empíricamente mediante la observación directa de su impacto en la calidad visual (ver Figura 2), influye directamente en la imagen resultante.

- **Ventana de trabajo:** Define el intervalo temporal utilizado para generar cada imagen, establecido en 30 segundos en los experimentos realizados.
- **Paso:** Representa el desplazamiento temporal entre imágenes consecutivas, fijado experimentalmente en 3 segundos.
- **Escalado:**Para asegurar que los valores de las imágenes se encuentren en el rango [0, 255], se consideraron métodos de normalización. Una métrica global utiliza toda la señal para normalizar, preservando la escala general. Una métrica local se basa en una ventana deslizante, mejorando el contraste. Finalmente, se adoptó una métrica combinada (ϵ=0.5) para equilibrar estas dos aproximaciones, tal como se define en la Ecuación (2).

$$\overline{x}^i = \epsilon(\overline{x}^i_{metrica-global}) + (1-\epsilon)(\overline{x}^i_{metrica-local}) \quad (2)$$

- **Saturación:** Para mitigar el ruido y valores atípicos, la señal se satura, asignando el valor 255 a los valores superiores al percentil 95 y el valor 0 a los inferiores al percentil 5.

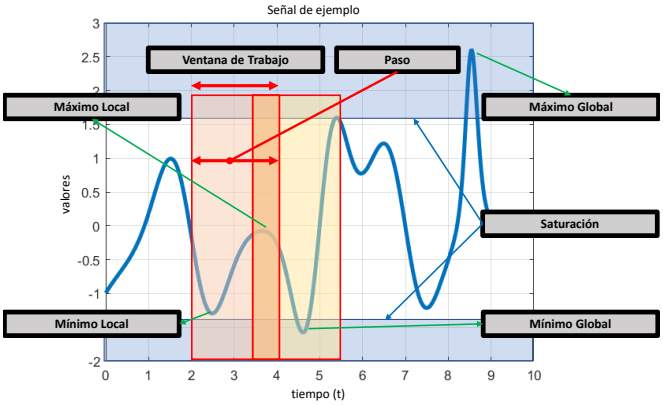

Figura 2: Representación visual de los parámetros durante la generación de imágenes a partir de señales fisiológicas: máximos y mínimos locales/globales, valores de saturación, ancho de ventana de trabajo y paso entre imágenes consecutivas.

3.2. *Recurrence Plot*

Los Gráficos de Recurrencia (Recurrence Plots (RP) por sus siglas en inglés) son una herramienta utilizada para visualizar los patrones repetitivos de una trayectoria $\vec{x} \in \mathbb{R}^d$ en el espacio de fases (Marwan et al., 2007). La Ecuación (3) muestra cómo se genera una matriz (que posteriormente se visualiza como una imagen), donde N es el número de puntos medidos, ϵ es un umbral de distancia, Θ es la función de Heaviside (definida como $\Theta(x) = 0$ si $x < 0$, y $\Theta(x) = 1$ en caso contrario), y $\|\cdot\|$ representa la norma utilizada para medir la distancia. En el caso de señales unidimensionales, donde $\vec{x} \in \mathbb{R}$, se tiene que $d = 1$, y la matriz RP se construye calculando diferencias

término a término entre los valores de la señal. El parámetro ϵ se utiliza para generar una imagen binaria (RP-binario), donde los elementos más cercanos se asignan con valor 1. Existe también la posibilidad de eliminar el uso de ϵ y de la función Θ, lo que da lugar a un gráfico de recurrencia continuo (RP-continuo) en lugar de binario.

$$R_{i,j}(\epsilon) = \Theta(\epsilon - \|\vec{x_i} - \vec{x_j}\|), \qquad i, j = 1, ..., N \tag{3}$$

3.3. Gramian Angular Field

El Campo Angular Gramiano (Gramian Angular Field, o GAF por sus siglas en inglés) es una técnica que transforma una serie temporal unidimensional en una imagen bidimensional. Primero se normalizan los datos (habitualmente en el rango [0,1]) y luego se mapean a coordenadas polares (Xu et al., 2020). A partir de los ángulos obtenidos, se construye una matriz utilizando funciones trigonométricas, lo que da lugar a dos variantes: el Campo Angular Gramiano de Suma (Gramian Angular Summation Field, o GASF) y el Campo Angular Gramiano de Diferencia (Gramian Angular Difference Field, o GADF). Estas representaciones permiten capturar correlaciones temporales y patrones en la serie original.

A continuación, se describe el procedimiento general del algoritmo, suponiendo que la serie X contiene N elementos, es decir, $X = [x_1, x_2, \ldots, x_N]$ (Xu et al., 2020):

1. Normalización de los valores a un rango adecuado para aplicar funciones trigonométricas. Existen dos opciones:
   - Normalización en el intervalo $[-1, 1]$:
   $$\overline{x}^i_{-1} = \frac{(x_i - \text{máx}(X)) + (x_i - \text{mín}(X))}{\text{máx}(X) - \text{mín}(X)} \tag{4}$$
   - Normalización en el intervalo $[0, 1]$:
   $$\overline{x}^i_{0} = \frac{x_i - \text{mín}(X)}{\text{máx}(X) - \text{mín}(X)} \tag{5}$$
2. Conversión de la serie temporal desde coordenadas cartesianas a coordenadas polares:
$$\begin{cases} \phi_i = \arccos(\overline{x}_i), & \text{con } \overline{x}_i \in [-1, 1] \\ r_i = \frac{i}{N}, & i = 1, \ldots, N \end{cases} \tag{6}$$
3. Cálculo del campo GAF en sus dos variantes:
   - **GASF (Gramian Angular Summation Field)**:
   $$\text{GASF} = \begin{bmatrix} \cos(\phi_1 + \phi_1) & \ldots & \cos(\phi_1 + \phi_N) \\ \cos(\phi_2 + \phi_1) & \ldots & \cos(\phi_2 + \phi_N) \\ \vdots & \ddots & \vdots \\ \cos(\phi_N + \phi_1) & \ldots & \cos(\phi_N + \phi_N) \end{bmatrix} \tag{7}$$
   - **GADF (Gramian Angular Difference Field)**:
   $$\text{GADF} = \begin{bmatrix} \sin(\phi_1 - \phi_1) & \ldots & \sin(\phi_1 - \phi_N) \\ \sin(\phi_2 - \phi_1) & \ldots & \sin(\phi_2 - \phi_N) \\ \vdots & \ddots & \vdots \\ \sin(\phi_N - \phi_1) & \ldots & \sin(\phi_N - \phi_N) \end{bmatrix} \tag{8}$$

3.4. Markov Transition Field

El Campo de Transición de Markov (Markov Transition Field o MTF por sus siglas en inglés) genera una matriz de transición de estados a partir de una serie temporal unidimensional. Para ello, se definen estados discretos en la serie y se calculan las probabilidades de transición de primer orden, lo que da lugar a una matriz de transición W. Esta matriz se proyecta de nuevo sobre los índices temporales de la serie original, formando así una representación bidimensional en la que cada elemento indica la probabilidad de transición entre estados a lo largo del tiempo (Yan et al., 2022).

A diferencia de otros métodos, el MTF se basa en el uso de probabilidades para transformar señales en imágenes. El algoritmo consiste en : Dada una serie temporal de la forma $X = [x_1, x_2, \ldots, x_N]$, y un conjunto de estados discretos definidos como $Q = [q_1, q_2, \ldots, q_Q]$, donde cada punto x_i de la serie puede asociarse a un estado q_j. El algoritmo de implementación consta de los siguientes pasos:

1. Calcular las transiciones entre los estados Q mediante una cadena de Markov de primer orden. Se genera una Matriz de Transición $W \in \mathbb{R}^{Q \times Q}$, donde cada elemento w_{ij} representa la probabilidad de que un estado q_j sea seguido por un estado q_i.
2. Normalizar la Matriz de Transición W para que cada fila sume 1.
3. Proyectar la Matriz de Transición W sobre los índices temporales de la serie original para generar el MTF, que es la imagen resultante $M \in \mathbb{R}^{N \times N}$. Esta imagen se escala a la dimensión adecuada si es necesario. Esto se representa mediante la ecuación (9).

$$M_{ij} = \begin{bmatrix} w_{ij}|x_1 \in q_i, x_1 \in q_j & \ldots & w_{ij}|x_1 \in q_i, x_n \in q_j \\ w_{ij}|x_2 \in q_i, x_1 \in q_j & \ldots & w_{ij}|x_2 \in q_i, x_n \in q_j \\ \vdots & \ddots & \vdots \\ w_{ij}|x_n \in q_i, x_1 \in q_j & \ldots & w_{ij}|x_n \in q_i, x_n \in q_j \end{bmatrix} \tag{9}$$

En el análisis realizado, se evaluaron dos configuraciones: una utilizando 4 estados (MTF-4) y otra con 128 estados (MTF-128), lo cual permite capturar la dinámica probabilística de la serie con distintos niveles de granularidad.

4. Arquitectura de fusión multimodal

La arquitectura de fusión multimodal propuesta, ilustrada en la Figura 3, permite integrar señales fisiológicas con información visual obtenida a partir de imágenes del conductor. Para hacer esto posible, las señales fisiológicas son previamente transformadas en representaciones bidimensionales utilizando los métodos del estado del arte descritos en la Sección 3. Esta conversión no solo incrementa la dimensionalidad de las señales, haciéndolas compatibles con modelos basados en visión por computador, sino que también preserva su estructura temporal, lo que resulta fundamental para tareas de análisis dinámico.

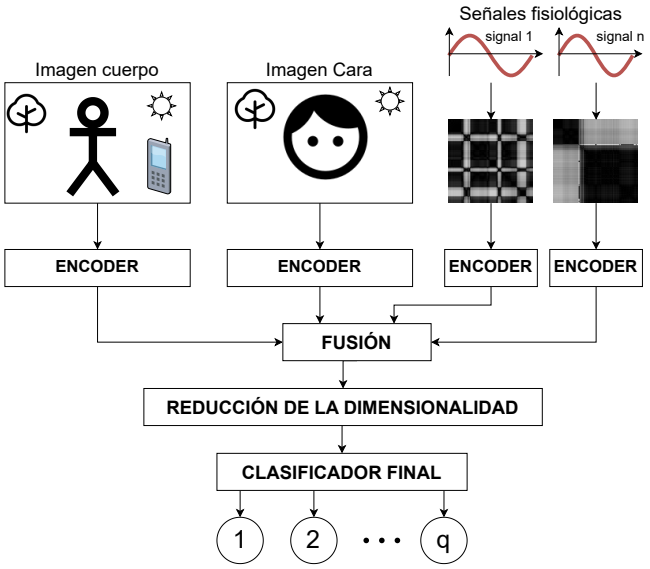

Figura 3: Arquitectura de fusión multimodal que integra señales fisiológicas transformadas en imágenes e imágenes de vídeo para clasificar el estado del conductor.

En cuanto a la componente visual, se emplean imágenes obtenidas de vídeo de la cara y del cuerpo del conductor. Se descarta el uso directo de autocodificadores de vídeo debido a las enormes exigencias computacionales asociadas al tratamiento de tensores con múltiples dimensiones. La Ecuación (10) muestra la estructura de un tensor de vídeo, donde B es el tamaño del lote (batch size), S el número de cámaras, T la duración en segundos, F los fotogramas por segundo, C el número de canales, H la altura y W el ancho de cada imagen. La alta dimensionalidad de este tensor genera dos problemas grandes para el entrenamiento: por un lado, la necesidad de disponer de un conjunto de datos de entrenamiento extremadamente grande; y por otro, el entrenamiento de modelos de gran escala con funciones objetivo altamente complejas. Para reducir esta complejidad, se adopta un enfoque más eficiente basado en el procesamiento de vídeo como secuencias de fotogramas (imágenes). Al tratar el vídeo como una serie de imágenes individuales, se elimina la dimensión temporal conjunta ($T \times F$), lo que simplifica significativamente el problema.

$$\text{Tensor de entrada: } \mathcal{X} \in \mathbb{R}^{B\times S\times T\times F\times C\times H\times W} \tag{10}$$

No obstante, dado que la arquitectura propuesta no incorpora explícitamente mecanismos para modelar la dependencia temporal entre los fotogramas, se propone como solución complementaria la aplicación de un Filtro Bayesiano Dinámico en el espacio latente de características. Los Filtros Bayesianos permiten estimar la evolución temporal de las representaciones, introduciendo información secuencial (Slavic et al., 2023). Esta estrategia permite incorporar dicha información sin aumentar la complejidad de la red principal.

La arquitectura propuesta combina fases de procesamiento supervisadas y no supervisadas. La reducción de dimensionalidad de las imágenes y las señales fisiológicas transformadas se realiza mediante autocodificadores entrenados sin etiquetas. Posteriormente, se lleva a cabo la fusión multimodal en un nivel intermedio. Si bien la fusión puede implementarse de diversas maneras, una opción es aplicar una operación de suma vectorial, seguida de una técnica de reducción de dimensionalidad no supervisada como el Análisis de Componentes Principales (PCA) para simplificar el espacio de características. Finalmente, se realiza una etapa de clasificación supervisada, que requiere un conjunto de datos etiquetado para estimar el estado del conductor en un conjunto definido de clases.

5. Resultados

5.1. *Comparativa Conversión de Señal a Imagen*

La arquitectura de fusión multimodal de la Figura 3 se ha adaptado en la mostrada en la Figura 4, la cual incluye la fusión de señales fisiológicas capturadas usando la pulsera Empatica E4 en entornos reales de conducción (Puertas-Ramirez et al., 2023).

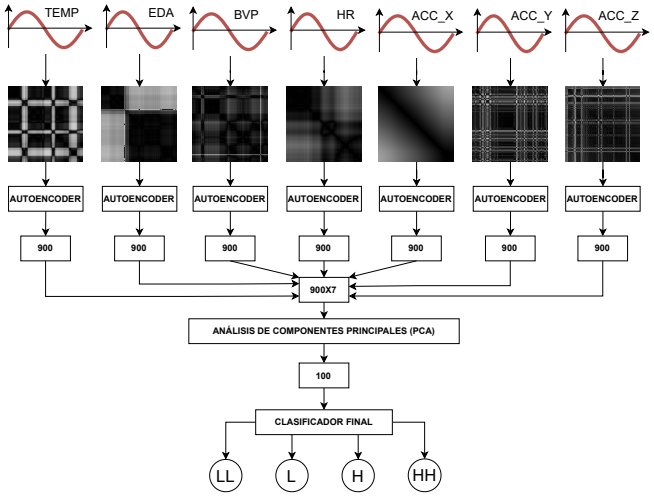

Figura 4: Adaptación de la arquitectura de la Figura 3 para la fusión multimodal de 7 señales fisiológicas.

Las señales empleadas son: aceleración en los tres ejes (ACCx, ACCy,ACCz), actividad electrodérmica de la piel (EDA), frecuencia cardíaca (HR), volumen de sangre en pulso (BVP) y temperatura corporal (TEMP). Todas las señales se han remuestreado a 4Hz y se han generado imágenes teniendo en cuenta la ventana de trabajo de 30 segundos usando las técnicas previamente mencionadas en la Sección 3. Cada imagen se pasa por el codificador para reducir su dimensionalidad, y los vectores resultantes se fusionan mediante concatenación. Para la posterior reducción de dimensionalidad del vector fusionado, se aplica PCA, reduciendo la dimensionalidad de ($7 \times 900 = 6300$) a 100 componentes. Finalmente, se emplea un clasificador de dos capas totalmente conectadas que clasifica el estado de atención del conductor en cuatro niveles: muy bajo (LL, Low-Low), bajo (L, Low), alto (H, High) y muy alto (HH, High-High). Su arquitectura simple y de entrenamiento rápido facilita el uso de datos no etiquetados y mejora la adaptabilidad a nuevos conductores.

Entre las transformaciones probadas, RP-continuo presenta la mayor precisión alcanzando un 70.3 %. Por el contrario, la RP-binario mostró un descenso significativo del rendimiento, con una precisión del 61,9 %. GASF alcanzó una precisión del 60,11 %, mientras que el GADF obtuvo la precisión

más baja, del 49,70 %. MTF-4 proporcionó una precisión del 63,39 % , pero descendió al 58,9 % cuando se utilizaron 128 estados (MTF-128). En general, la RP-continuo proporcionó sistemáticamente los mejores resultados y se seleccionó como método preferido para convertir las señales en imágenes durante el resto del estudio. Cabe destacar que hubo una diferencia del 20 % en la precisión entre los métodos con mejores y peores resultados, lo que subraya la importancia de evaluar múltiples técnicas de transformación para determinar el enfoque más eficaz.

5.2. *Autocodificadores en Imágenes*

Para el tratamiento de imágenes se prueba con los datos del *State-farm-distracted-driver-detection*(Montoya et al., 2016) que contiene 10 clases diferentes de acciones que puede realizar el conductor de un vehículo. Originalmente, se divide en 22.424 imágenes para entrenamiento y 79.726 imágenes para validación. Sin embargo, debido a la naturaleza de los autocodificadores, en este trabajo se ha invertido esta distribución, utilizando el conjunto más pequeño para validación y el más grande para entrenamiento. Se han evaluado diferentes Funciones de Pérdida Perceptual, y la Figura 5 muestra los resultados de emplear YOLOv8, VGG16, y DeepLabv3-ResNet50, comparados con la arquitectura entrenada con la pérdida MSE (Error Cuadrático Medio) y la imagen original.

Figura 5: Comparativa de de diferentes Funciones de Pérdida Perceptual en autocodificadores.

6. Conclusiones y Trabajos futuros

La arquitectura de fusión multimodal descrita en este artículo ha demostrado ser versátil para la fusión de información proveniente de diferentes sensores. El uso de RP-continuo, permite no solo aumentar la dimensionalidad de las señales fisiológicas, sino también integrar su inherente carácter temporal de una manera visual que facilita su procesamiento posterior por redes neuronales avanzadas. En el tratamiento de imágenes, las Funciones de Pérdida Perceptual pueden mejorar la representación aprendida por el modelo, aunque la selección de la función óptima es un desafío. Es importante notar que, en la reducción de dimensionalidad con autocodificadores, el objetivo principal es obtener una representación útil en el espacio latente, más que la reconstrucción visual perfecta.

En el futuro, se continuará este trabajo centrando los esfuerzos en el tratamiento detallado de imágenes del cuerpo y la cara del conductor, con el objetivo de predecir con mayor precisión su estado dentro de un vehículo autónomo. Además, se aplicará la arquitectura de fusión multimodal desarrollada para integrar de manera óptima señales fisiológicas e información visual.

Agradecimientos

Este trabajo ha sido realizado parcialmente gracias al apoyo de las subvenciones PID2021-124335OB-C21, PID2022-140554OB-C32, TED2021-129485B-C44, los proyectos TEC-2024/ECO-277 y TEC-2024/TEC-102 financiados por la Comunidad de Madrid y la ayuda de Formación de Profesorado Universitario (FPU-2023) financiada por el Ministerio de Ciencia, Innovación y Universidades del Gobierno de España.

Referencias

Bank, D., Koenigstein, N., Giryes, R., 2023. Autoencoders. Springer International Publishing, Cham. pp. 353–374. doi:10.1007/978-3-031-24628-9_16.

Deng, M., Gluck, A., Zhao, Y., Li, D., Menassa, C.C., Kamat, V.R., Brinkley, J., 2024. An analysis of physiological responses as indicators of driver takeover readiness in conditionally automated driving. Accident Analysis & Prevention 195, 107372. doi:https://doi.org/10.1016/j.aap.2023.107372.

Kazemi, M., Rezaei, M., Azarmi, M., 2025. Evaluating driver readiness in conditionally automated vehicles from eye-tracking data and head pose. IET Intelligent Transport Systems 19, e70006. doi:https://doi.org/10.1049/itr2.70006.

Marcano, M., Díaz, S., Pérez, J., Irigoyen, E., 2020. A review of shared control for automated vehicles: Theory and applications. IEEE Transactions on Human-Machine Systems 50, 475–491. doi:10.1109/THMS.2020.3017748.

Marwan, N., Carmen Romano, M., Thiel, M., Kurths, J., 2007. Recurrence plots for the analysis of complex systems. Physics Reports 438, 237–329. doi:https://doi.org/10.1016/j.physrep.2006.11.001.

Montoya, A., Holman, D., SF_data_science, Smith, T., Kan, W., 2016. State farm distracted driver detection. https://kaggle.com/competitions/state-farm-distracted-driver-detection. Kaggle.

Pihlgren, G.G., Sandin, F., Liwicki, M., 2020. Improving image autoencoder embeddings with perceptual loss, in: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–7. doi:10.1109/IJCNN48605.2020.9207431.

Puertas-Ramirez, D., Fernandez-Matellan, R., Martin-Gomez, D., G. Boticario, J., Tena-Gago, D., 2023. Improving Autonomous Vehicle Automation Through Human-System Interaction. The 37th annual European Simulation and Modelling Conference , 294–300.

SAE International, 2021. Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles. Technical Report. SAE International. doi:10.4271/J3016_202104.

Slavic, G., Alemaw, A.S., Marcenaro, L., Martín Gómez, D., Regazzoni, C., 2023. A kalman variational autoencoder model assisted by odometric clustering for video frame prediction and anomaly detection. IEEE Transactions on Image Processing 32, 415–429. doi:10.1109/TIP.2022.3229620.

Wang, J., Yang, X., Wang, Z., Wei, X., Wang, A., He, D., Wu, K., 2024. Efficient mixture-of-expert for video-based driver state and physiological multi-task estimation in conditional autonomous driving. arXiv preprint arXiv:2410.21086 doi:https://doi.org/10.48550/arXiv.2410.21086.

Weigl, K., Schartmüller, C., and, A.R., 2023. Development of the questionnaire on non-driving related tasks (qndrt) in automated driving: revealing age and gender differences. Behaviour & Information Technology 42, 1374–1388. doi:10.1080/0144929X.2022.2073473.

Xu, H., Li, J., Yuan, H., Liu, Q., Fan, S., Li, T., Sun, X., 2020. Human activity recognition based on gramian angular field and deep convolutional neural network. IEEE Access 8, 199393–199405. doi:10.1109/ACCESS.2020.3032699.

Yan, J., Kan, J., Luo, H., 2022. Rolling bearing fault diagnosis based on markov transition field and residual network. Sensors 22. doi:10.3390/s22103936.