Especificación Técnica: Monitoreo de Atención de Próxima Generación

Arquitectura Híbrida MediaPipe + L2CS-Net

Reporte de Investigación y Guía de Implementación

27 de enero de 2026

Introducción

El monitoreo de la atención estudiantil en entornos virtuales requiere una precisión que supere la simple detección de presencia. La limitación principal de los sistemas actuales es la incapacidad de detectar el desvío de la mirada (*gaze*) cuando la cabeza permanece estática. Esta especificación propone una solución basada en modelos de apariencia, integrando estimación de pose 6-DoF y vectores de mirada tridimensionales.

Fundamentos Matemáticos

2.1. Estimación de Pose de la Cabeza (HPE)

Para determinar la orientación de la cabeza, se utiliza el algoritmo **Perspective-n-Point (PnP)**, que resuelve la relación entre puntos 3D de un modelo facial genérico y sus proyecciones 2D en la imagen:

$$s \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \mathbf{A} \begin{bmatrix} \mathbf{R} & | & \mathbf{t} \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \tag{1}$$

Donde:

- **A** es la matriz intrínseca de la cámara.
- **R** es la matriz de rotación, de la cual se extraen los ángulos de **Yaw** (γ), **Pitch** (ρ) y **Roll** (ϕ).
- Umbrales críticos de distracción: $|\gamma| > 25°$ o $|\rho| > 15°$.

2.2. Mirada Basada en Apariencia (L2CS-Net)

A diferencia de los métodos geométricos, L2CS-Net utiliza una CNN para predecir los ángulos de la mirada directamente desde la imagen (*Appearance-based gaze estimation*). Los ángulos de salida son:

- **Gaze Pitch** (α): Movimiento vertical del ojo.
- **Gaze Yaw** (β): Movimiento horizontal del ojo (ideal para detectar desvíos laterales sin mover la cabeza).

2.3. Detección de Parpadeo (EAR)

El *Eye Aspect Ratio* permite cuantificar la apertura palpebral:

$$EAR = \frac{\|p_2 - p_6\| + \|p_3 - p_5\|}{2\|p_1 - p_4\|} \tag{2}$$

Un estado de distracción o fatiga se detecta cuando el promedio de $EAR < 0{,}25$ por un tiempo prolongado.

Arquitectura del Sistema Propuesto

Componente	Tecnología	Responsabilidad
Frontend	Next.js / MediaPipe	Landmark detection (468 pts) y UI.
Backend	FastAPI / PyTorch	Inferencia de L2CS-Net y SolvePnP.
Comunicación	WebSockets	Streaming de frames a 15-20 FPS.
Filtrado	Filtro de Kalman	Suavizado de jitter en mirada y pose.

Cuadro 1: Distribución de la lógica del sistema.

Índice de Compromiso (Engagement Index)

Se define el índice EI como una función ponderada de los estados detectados:

$$EI(t) = w_g \cdot GazeScore(t) + w_p \cdot PoseScore(t) + w_b \cdot BlinkScore(t) \tag{3}$$

Pesos sugeridos: $w_g = 0{,}5, w_p = 0{,}3, w_b = 0{,}2$.