Documentacion en sucio   
  
Investigación

**Técnicas de *Data Augmentation* para lectura labial**

**OpenCV + Albumentations**

**Al usar Albumentations.ReplayCompose, se generan aumentaciones aleatorias basadas en un conjunto de técnicas predefinidas. En tu script, especificaste un conjunto de transformaciones posibles, pero no se aplican todas a cada vídeo, sino que se elige aleatoriamente en cada ejecución de aumento.**

**🎨 Técnicas que definiste (las posibles en tu código)**

**Estas son las 7 transformaciones que el sistema puede aplicar, en distintas combinaciones:**

| **Técnica** | **Descripción** |
| --- | --- |
| **HorizontalFlip** | **Invierte la imagen horizontalmente (como un espejo).** |
| **RandomBrightnessContrast** | **Ajusta brillo y contraste de forma aleatoria.** |
| **ShiftScaleRotate** | **Desplaza, escala y rota la imagen levemente.** |
| **GaussianBlur** | **Aplica desenfoque gaussiano.** |
| **MotionBlur** | **Simula desenfoque por movimiento.** |
| **RandomGamma** | **Ajusta la corrección gamma (afecta brillo general).** |
| **ImageCompression** | **Reduce calidad de compresión JPEG (simula pérdida).** |

**💡 ¿Qué combinación se usó en cada vídeo?**

**Como usaste ReplayCompose, el sistema guardó internamente qué transformaciones se aplicaron a cada vídeo.**

**Pero: no lo estás guardando tú explícitamente. Por eso, aunque hayas creado 500 vídeos aumentados, ahora mismo no puedes saber exactamente qué combinación se usó en cada uno... 🧠**

**📄 Estudio 1: "A Lip-Reading Model for Tagalog Using Multimodal Deep Learning Approach"**

**Objetivo principal**

Desarrollar un modelo de lectura de labios específico para el idioma tagalo utilizando una arquitectura de aprendizaje profundo multimodal. El modelo busca mejorar la comunicación entre hablantes nativos y no nativos sordos o con problemas auditivos.

**Enfoque multimodal**

* El modelo se basa en una **combinación de datos visuales (movimientos de labios)** y **contexto lingüístico (texto)**.
* Se inspira en la arquitectura **LipNet**, pero sin depender del componente auditivo, lo que lo hace más accesible.
* Emplea una **combinación de redes neuronales convolucionales (CNN)** para la parte visual, y **Long Short-Term Memory (LSTM)** para el análisis de secuencias en el lenguaje.

**Metodología**

* Se diseña un modelo híbrido CNN-LSTM:
  + **CNN** se encarga de la detección de la región de los labios y el procesamiento visual.
  + **LSTM** interpreta la secuencia temporal de los movimientos de labios y contextualiza con el lenguaje.
* El preprocesamiento es clave: los videos son divididos en 75 fotogramas, se extraen características faciales, se enfocan en los labios, se convierten a escala de grises y se normalizan.
* Todo esto se combina con información textual convertida en tokens numéricos.

**Resultados**

* **Precisión de validación de 89.5%**.
* **Velocidad de procesamiento aumentó en un 25%**, eliminando la necesidad de datos de audio.
* Se alcanzaron tasas de verdaderos positivos y negativos del 91% y 87% respectivamente.
* El modelo demostró ser escalable, robusto y adaptable a diferentes estilos de habla y acentos.

**Conclusiones y recomendaciones**

* La combinación de información visual y textual es altamente efectiva.
* Recomiendan ampliar este enfoque a otros idiomas poco representados.
* Subrayan la importancia de adaptar la arquitectura a las características fonéticas del idioma tratado

**🤖 Estudio 2: "Integrating Persian Lip Reading in Surena-V Humanoid Robot for Human-Robot Interaction"**

**Objetivo principal**

Incorporar la tecnología de lectura de labios al robot humanoide **Surena-V**, permitiéndole comprender palabras comunes en persa en entornos ruidosos o sin audio.

**Dataset propio**

* 20 participantes grabados pronunciando 7 palabras persas comunes como "Hola", "Ven", "Toma", "Escribe", etc.
* Capturado con cámara RGB-D para obtener información tanto visual como de profundidad.
* Las grabaciones fueron recortadas en segmentos individuales de 1 segundo (20 frames), resultando en **2800 clips de video** procesados para entrenamiento.

**Métodos explorados**

1. **Método indirecto**:
   * Basado en el seguimiento de **puntos de referencia faciales** (landmarks) alrededor de los labios.
   * Se normalizan los datos para reducir el efecto de la luz, ángulos de cámara o movimientos menores de cabeza.
   * Modelos como VGG19, ResNet y **MobileNet** fueron probados.
   * MobileNet logró la mejor precisión en este enfoque: **52%**.
2. **Método directo**:
   * Analiza directamente los datos crudos de video utilizando:
     + **CNN**: buen rendimiento general, alcanzando **75%** de precisión.
     + **LSTM**: sobresaliente en precisión (**89%**), reconociendo mejor las secuencias temporales y comandos verbales.
   * El modelo LSTM fue el elegido para **implementación en tiempo real** en el robot Surena-V.

**Implementación real**

* El robot fue capaz de **reconocer palabras en tiempo real** y reaccionar adecuadamente (gestos, acciones, comandos).
* La integración con un detector de objetos (YOLOv5) permitió responder de forma más completa.

**Limitaciones y futuras mejoras**

* Problemas con movimientos bruscos o ciertas distancias en el método indirecto.
* Se sugiere el uso de técnicas como **"optical flow"** para capturar mejor los movimientos finos de los labios.
* Ampliar el vocabulario y diversidad del dataset para mejorar la generalización.

Problemas

A green light on a person's face

AI-generated content may be incorrect.