



UNCUYO
UNIVERSIDAD
NACIONAL DE CUYO



FACULTAD
DE INGENIERÍA

Trabajo final Inteligencia Artificial 1 (2024)

Visión artificial y reconocimiento de voz

Universidad Nacional de Cuyo, Facultad de Ingeniería

Autor: Francisco Castel

2024/2025

Introducción

Visión Artificial

- Identificación de objetos mediante imágenes.
- Aplicaciones: automatización, robótica, diagnóstico médico.

Reconocimiento de Voz

- Conversión de audio en texto o comandos.
- Uso de coeficientes cepstrales y modelos de clasificación.

Desafíos abordados

- Reducir dimensionalidad.
- Clasificación robusta en condiciones controladas.

Resumen

Objetivo

- Combinar visión artificial y reconocimiento de voz para clasificar verduras, en concreto.
 - Berenjena
 - Camote
 - Papa
 - Zanahoria

Técnicas utilizadas

- KMeans para imágenes
- KNN para audios

Enfoque

- Extracción de características para cada tipo de dato
- Reducción de dimensionalidad para optimización y coherencia

Clasificador de Imágenes (No Supervisado)

- Algoritmo: **KMeans**.
- Características:
 - Momentos de Hu.
 - Color promedio.
- Entorno controlado:
 - Fondo constante.
 - Iluminación uniforme.

Clasificador de Audio (Supervisado)

- Algoritmo: **KNN**.
- Características:
 - MFCCs (media, máximo, mínimo, desviación estándar).
 - Espectro de Potencia
 - Entropía energética
 - RMS y duración.
- Entorno controlado:
 - Bajo nivel de ruido.

Agente	Rendimiento	Entorno	Actuadores	Sensores
Clasificador de imágenes con aprendizaje no supervisado	Precisión: Si la predicción de la verdura es correcta (es decir, la predicción es igual a la categoría real de la verdura).	Entorno controlado con iluminación neutra. Las imágenes se toman siempre con el mismo fondo, y la rotación de las verduras no afecta a la clasificación.	Pantalla de la computadora que muestra la predicción de la verdura al usuario. No hay acción física directa sobre el entorno.	Cámara fotográfica de celular con flash (captura imágenes en formato RGB).
Clasificador de audio, aprendizaje supervisado	Precisión: La precisión con la que el sistema clasifica los audios correctamente, es decir, si la predicción de la palabra es correcta.	El entorno es controlado, ya que el sistema funciona en condiciones de bajo nivel de ruido. No es adecuado para ambientes con alto nivel de ruido.	Pantalla de la computadora que muestra la predicción del audio al usuario. No hay interacción física directa con el entorno.	Micrófono y driver de audio con al menos una tasa de muestreo de 16KHz.

Cuadro: Tabla REAS de ambos agentes

Propiedades del entorno

Aunque se traten como agentes diferentes, pueden condensarse las propiedades a un entorno compartido ya que presenta propiedades iguales.

- Determinístico
- Totalmente observable
- Secuencial
- Estático
- Discreto
- Mono agente

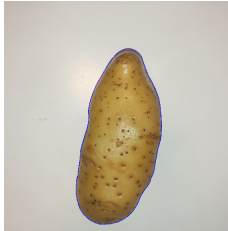
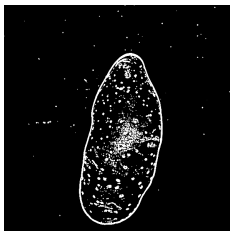
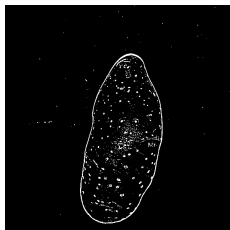
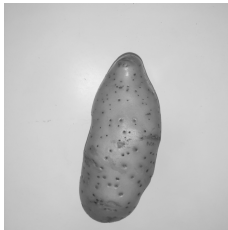
Preprocesamiento de imágenes

- Pasos

- 1 Desenfoque gaussiano
- 2 Escala de grises
- 3 Umbralización adaptativa
- 4 Operación morfológica

- Resultados

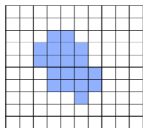
- Mejora en la detección de contornos
- Reducción de ruido sal y pimienta
- Características finales: Momentos de Hu y color promedio



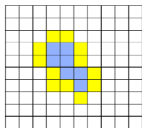
Filtro morfológico

Consiste en dos operaciones:

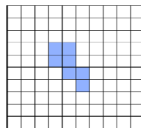
Erosión



(a)

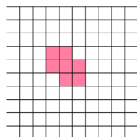


(b)

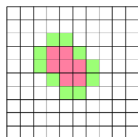


(c)

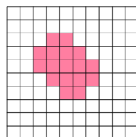
Dilatación



(a)



(b)



(c)

A partir de ellas se pueden lograr dos operaciones compuestas muy interesantes.


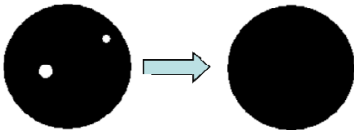
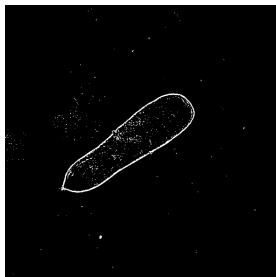
Apertura $I \circ H = (I \ominus H) \oplus H$	Cierre $I \bullet H = (I \oplus H) \ominus H$
	

Figura 3.- Ejemplos de operaciones de apertura (izquierda) y cierre (derecha).

Binarización adaptativa

La binarización adaptativa es una solución práctica al inconveniente de la selección de un umbral fijo para la totalidad de la imagen.

En lugar de tener un único umbral como la media de los dos valores predominantes, la media se calcula de forma local mediante una ventana definida.



Preprocesamiento de audios

Pasos:

- Pre-énfasis: Resalta altas frecuencias.
- Filtro paso banda: Elimina ruido fuera del rango útil.
- Reducción de ruido: Basada en la transformada de Fourier.
- Normalización de loudness: Ajusta volumen.
- División en 4 segmentos

Características finales por cada segmento:

- MFCCs (media, máximo, mínimo, desviación estándar).
- Espectro de potencia
- Entropía de energía
- RMS y duración total.

MFCC (Mel-Frequency Cepstral Coefficients)

Los MFCC son una representación compacta de las características acústicas de una señal de audio.

- Objetivo: Capturar la información espectral relevante para el habla.
- Proceso:
 - Transformada de Fourier para obtener el espectro de la señal.
 - Mapeo a una escala mel, que imita la percepción humana del sonido.
 - Aplicación de la transformada discreta de coseno (DCT) para reducir la dimensionalidad.

Espectro de Potencia, Entropía y RMS

- **Espectro de Potencia:** Representa cómo se distribuye la energía de la señal en diferentes frecuencias.
 - Muestra frecuencias predominantes (usado en análisis de voz e instrumentos).
- **Entropía de la Energía:** Mide el desorden en la distribución de energía a través de las frecuencias.
 - Alta entropía: energía distribuida de forma compleja.
 - Baja entropía: energía concentrada en pocas frecuencias.
- **RMS (Root Mean Square):** Mide la magnitud promedio de una señal de audio.
 - Representa la energía total de la señal.
 - Se calcula como la raíz cuadrada de la media de los cuadrados de los valores de la señal.

Reducción de dimensionalidad

Clasificador de imágenes: 5 a 3 dimensiones

Clasificador de audio: 209 a 3 dimensiones

A partir de ciertas pruebas se decide UMAP antes que PCA.
Se experimento con DBSCAN para la eliminación de muestras incoherentes.

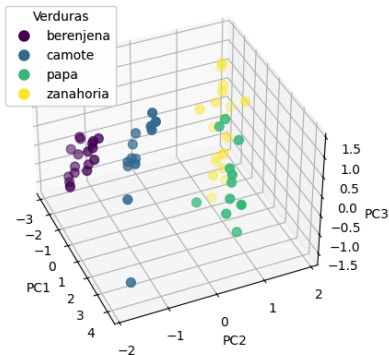
Solo para visualizar los datos?

Una buena práctica

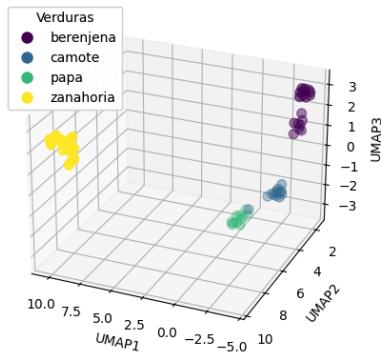
La reducción de dimensionalidad es fundamental en los algoritmos como KNN, esto es para poder hacer frente a la *maldición de la dimensionalidad* la cual hace que a medida que aumentemos las dimensiones de nuestros datos se rompan las condiciones para que KNN tenga sentido.

Datos redimensionados (Imágenes)

PCA



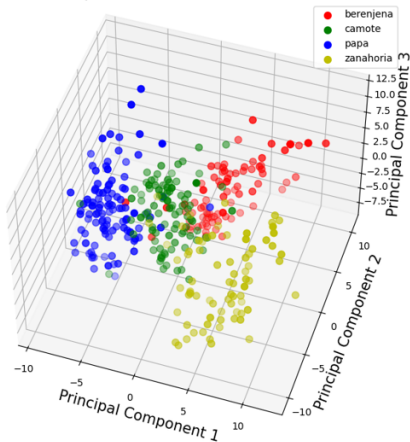
UMAP



Datos redimensionados (Audio)

PCA

3 Componentes PCA (Sin Outliers)



UMAP

Visualización UMAP 3D

