

UNIVERSIDAD DE LOS ANDES

DEPARTAMENTO DE FÍSICA



Proyecto computacional:

ESTUDIO DE ENTROPÍA Y COMPLEJIDAD EN IMÁGENES

Presentado por:

Karol Yessenia Rivera Rengifo

Asesor:

Jaime Forero Romero Ph.D.

29 de noviembre de 2021

1. Introducción

Gracias a los avances tecnológicos recientes, la información recolectada a partir del estudio de sistemas científicos suele ser robusta y requerir de varias etapas para su correcto análisis. El uso de imágenes como archivos con gran capacidad de compresión de información es cada vez más frecuente y hace evidente la necesidad de crear métodos que permitan cuantificar la cantidad de información codificada. Usualmente, el análisis de imágenes científicas se ha concentrado en estudios de intensidad, cuyos cálculos asociados con entropía y complejidad se reducen al análisis de esta variable. Sin embargo, desde el establecimiento de nuevas definiciones de entropía y complejidad enmarcadas en la teoría de la información, se han considerado nuevos mecanismos para estudiar la composición de imágenes. Lo relevante de la nueva definición de dichas cantidades radica en que la teoría de la información es una rama de la teoría de la probabilidad que se encarga de estudiar las leyes matemáticas que rigen la transmisión y procesamiento de la información. De esta forma, el análisis que se le realice a las imágenes puede relacionarse con la cantidad de información contenida al calcular su entropía. Recientemente se ha realizado un estudio a obras de arte donde se aplicó el concepto de entropía y complejidad utilizando permutaciones en lugar del típico análisis de intensidades. En este trabajo se encontró que las imágenes tenían valores de entropía y complejidad similares si pertenecían a épocas similares. En este contexto valdría la pena preguntarse si el método para el cálculo de entropía y complejidad para obras de arte puede ser utilizado también en imágenes provenientes de estudios científicos, y si permitiría asociar valores de entropía y complejidad a imágenes que compartan características en común.

Este proyecto busca entonces aplicar un método novedoso para cálculo de entropía y complejidad a imágenes bidimensionales. Por lo tanto, es necesario construir un banco de imágenes bidimensionales que provengan de ciencia, en particular de astronomía y microscopía; y otras adicionales con el fin de observar el comportamiento del plano entropía y complejidad en el caso de rostros y obras de arte. Adicionalmente, se debe diseñar un programa que permita realizar el cálculo de entropía y complejidad utilizando como lenguaje de programación python. Y finalmente, se aplicará el código a las imágenes para obtener dispersiones en el plano entropía complejidad.

Estudiar el método para cálculo de entropía y complejidad basado en permutaciones es importante porque puede contribuir al desarrollo de métodos de análisis complementarios a los usuales basados en intensidad. Adicionalmente, el método es llamativo porque está

basado en el análisis de permutaciones de celdas, una variable diferente a la típicamente utilizada para estudios de entropía: la intensidad. Por otro lado, es propuesto para sistemas bidimensionales, los cuales no cuentan con el mismo nivel de profundización en lo que respecta a métodos para cálculos de entropía y complejidad para sistemas 1D, por ejemplo para señales. Finalmente, los resultados del estudio base abren la posibilidad de encontrar un método de clasificación de imágenes de acuerdo con sus valores de entropía y complejidad.

2. Estado del arte

El artículo publicado en 1948 por Claude Shannon de nombre *A mathematical theory of communication* representó un hito en la comprensión de qué es la información y cómo se relaciona con la aleatoriedad. En su artículo Shannon tenía como objetivo hacer la transmisión de información lo más eficiente posible, y para ello identificó los 4 aspectos limitantes: el ruido, el tamaño de la información, el ancho de banda y la velocidad de transmisión. Con el fin de optimizar el proceso de transmisión de información y de comprender mejor las variables limitantes, Shannon estudió 3 conceptos fundamentales: qué es la información y cómo se cuantifica, cuál es la capacidad de transmisión del medio, y cómo codificar la información para utilizar el canal de transmisión a su máxima capacidad (Holik, 2016). Shannon define la información como el grado de indeterminación de un estado (p representa la probabilidad del estado) (Shannon, 1948):

$$I = -\log(p), \quad (1)$$

mientras que la entropía la define como la información promedio ponderada contenida (Shannon, 1948):

$$H(p) = - \sum_{n=0}^c p_n \log(p_n) = \sum_{n=0}^c p_n I. \quad (2)$$

Una de las propiedades más importantes de la entropía de Shannon es que su valor oscila entre 0 y logaritmo del número de microestados disponibles. De esta forma, el valor de entropía es mínimo cuando un solo microestado cuenta con toda la probabilidad y máxima cuando todos los microestados tienen la misma probabilidad (Holik, 2016). También cabe mencionar que, en teoría de la información la base de logaritmo que se usa es el logaritmo

natural.

Por otra parte, Shannon establece que la entropía puede entenderse también cómo la compresión óptima de la información a transmitir (Holik, 2016 - primer teorema), y que mientras se mantenga la tasa de transmisión menor a la máxima del canal, la probabilidad de error no aumentará (Holik, 2016). En el trabajo la relación entre entropía e información se hace matemáticamente evidente, sin embargo, implica algo mucho más profundo: la información implica la existencia de la aleatoriedad. De esta manera, estados que cuentan con mayor probabilidad de ocurrencia aportarían menos información, mientras que los eventos poco probables serían más significativos (Shannon, 1948). Por otra parte, el trabajo de Shannon lleva al posterior desarrollo de la teoría de la información: una rama de la teoría de la probabilidad que estudia las leyes matemáticas que rigen la transmisión y procesamiento de la información (MacKay, 2003).

Como se mencionaba en un principio, el trabajo de Shannon iba orientado a la optimización del proceso de transmisión de información, es decir, estaba relacionado con el análisis de señales. En este contexto, el primer y segundo teorema fueron fundamentales para identificar los mejores métodos de codificación, los canales que tenían la mayor tasa de transmisión, y cómo reducir al máximo los errores en las señales (Holik, 2016). En cuanto al análisis de imágenes científicas, previo al trabajo de Shannon se utilizaba la definición usual de entropía proveniente de la física. La entropía se entendía como la cantidad de microestados que se ajustan a un macroestado de equilibrio y fue desarrollada conceptualmente por Clausius (1850) y matemáticamente por Boltzmann (1877). Con esta definición, las investigaciones se limitaban a estudiar una variable a la que se le asignaba un conjunto de microestados posibles para analizar un macroestado obtenido por medio de experimentos (Reif, 1998). La variable utilizada usualmente era la intensidad, puesto que se trata de una cantidad que puede obtenerse a partir del estudio de los píxeles que conforman las imágenes (Herrera, Murguía, y Quintana, 2018). Ejemplos de este tipo de trabajos son el descubrimiento de la estructura del ADN apoyado en patrones de difracción de rayos x (Fierro, 2001) o la identificación de los elementos que conforman las estrellas a partir de espectros, por ejemplo diferenciando entre metales y no metales (Espín, Toledo, Vázquez, Hernández, y Cielo, 2015). Estas técnicas han evolucionado para ser utilizadas en estudios más especializados, por ejemplo, para estudiar las condiciones físicas de envolturas gaseosas a partir de la emisión de líneas IR, así como la influencia del viento estelar en la emisión de dichas líneas (Granada, 2010).

Posterior al trabajo de Shannon se adoptó en algunas áreas la nueva definición de entropía

y se utilizó para el análisis de imágenes considerando variables similares a la intensidad o al color (Herrera, Murguía, y Quintana, 2018). En este contexto, los autores del artículo *History of art painting through the lens of entropy and complexity* proponen un nuevo método para calcular la entropía y complejidad de imágenes (Herrera y cols., 2018), el cual está basado en el estudio de permutaciones de una celda. Este método fue utilizado para estudiar obras de arte, encontrando que es útil para clasificarlas ya que la entropía de las imágenes suele relacionarse con la época a la que pertenecen (Sigaki, Perc, y Ribeiro, 2018). Además, mostró que la entropía obtenida podía compararse con la de una imagen que represente ruido, con el fin de identificar qué tanto difiere la obra de esta (Sigaki y cols., 2018). Este método también se ha aplicado al análisis de textura en imágenes digitales utilizadas normalmente en diseño (Antonelli, 2017). Con respecto a imágenes de ciencia, recientemente se ha propuesto un método basado en entropía de permutación para el análisis de imágenes texturizadas (Morel y Humeau-Heurtier, 2021). Entre las imágenes analizadas hay algunas provenientes de la medicina, las cuales son analizadas usualmente para determinar histopatologías. Los resultados muestran que el cálculo de complejidad basado en entropía de permutación permite diferenciar entre texturas y que además permite un procesamiento más eficiente en comparación con las mediciones usuales de complejidad (Morel y Humeau-Heurtier, 2021).

Por otra parte, las divergencias han jugado un papel fundamental en el análisis de imágenes ya que han sido utilizadas como una herramienta para cuantificar la complejidad, aunque también han sido relevantes en el estudio de series temporales, en el estudio de estados cuánticos, o series de símbolos (Bussandri, Linck, Ré, y Lamberti, 2014).

Las divergencias se entienden como una regla para establecer una comparación cuantitativa entre dos o más distribuciones de probabilidad. La divergencia de Jensen Shannon es un funcional proveniente de la teoría de la información y se utiliza para comparar la composición de los símbolos de dos secuencias en principio diferentes (Bussandri y cols., 2014). Considerando p_i la distribución de probabilidad para k símbolos, $p_i(e_k)$ la probabilidad del símbolo e_k y $\pi(i)$ el peso de la distribución i la divergencia de Jensen Shannon se define entonces como:

$$D(p_1, p_2) = H\left(\sum_{i=1}^2 \pi(i) p_i\right) - \sum_{i=1}^2 \pi(i) H(p_i), \quad (3)$$

siendo H la entropía de Shannon. Cabe mencionar que para que esta expresión sea válida se deben cumplir los siguientes dos requerimientos:

$$\sum p_i = 1, \quad \sum \pi(i) = 1. \quad (4)$$

Esta divergencia tiene la propiedad de que debe ser mayor a cero, a menos que ambas distribuciones de probabilidad sean las mismas, en cuyo caso la divergencia tomará un valor de cero (Lin, 1991). Adicionalmente, esta divergencia es simétrica con respecto al orden de las distribuciones de probabilidad.

Las divergencias tienen variadas interpretaciones dependiendo del área en el que se utilicen: 1) En el área de la física estadística la divergencia se entiende como una entropía de mezcla; 2) En la teoría de la información la divergencia se entiende como la información mutua, un tema muy estudiado por Shannon en su artículo de 1948 (Holik, 2016).

Como se mencionó anteriormente, la divergencia se considera una herramienta para definir la complejidad. En este caso se busca definir la complejidad estadística. La definición de complejidad estadística estará relacionada con la existencia de patrones, así tanto el ruido como un patrón periódico pueden considerarse poco complejos por tener, en el primer caso, un comportamiento muy simple, y el segundo caso, un comportamiento no trivial (San Martín, 2013). Con esta interpretación lo que se buscaría cuantificar con la complejidad es el grado de orden, dónde los extremos de orden y desorden serían un mínimo en la complejidad.

Debido a que las características del sistema se pueden ver reflejada en la distribución de probabilidad, se espera que la complejidad sea una función de dicha distribución. Por otra parte, dado que el conocimiento sobre el sistema es probabilístico, debe haber alguna influencia de la incerteza, es decir, debe involucrar una medida de información. Para finalizar, se debería poder medir qué tan separada se encuentra una distribución que represente la máxima incerteza (San Martín, 2013). Es por esta razón que la complejidad se define en términos de la divergencia de Jensen Shannon que busca cuantificar qué tan diferente es una distribución de probabilidad relativa a la distribución uniforme (en este caso sería quien representaría la máxima incerteza como se mencionó cuando se habló de las propiedades de la entropía de Shannon) y que es además una función de la entropía de Shannon, que como se vio previamente, es una función de la distribución de probabilidad y puede entenderse como una ponderación de la información contenida. La complejidad también dependerá de una constante de normalización D^* que será el valor máximo de la divergencia de Jensen Shannon.

$$C(p) = \frac{D(p, U)H(p)}{D*}. \quad (5)$$

3. Metodología

El proyecto se desarrolló atendiendo al cumplimiento de 3 fases: investigación, implementación y aplicación.

En la fase de investigación se estudió la teoría básica que permitiría el desarrollo del algoritmo computacional. Inicialmente se revisaron los fundamentos teóricos sobre entropía y complejidad en el marco de la teoría de la información y el análisis de imágenes. A continuación se investigó sobre los métodos previos para calcular entropía y complejidad en el contexto científico. Finalmente, se estudió la metodología basada en permutaciones propuesta en el artículo guía del proyecto. En esta fase se revisaron principalmente artículos y libros de texto, los cuales se encuentran debidamente referenciados al final del presente documento, mientras que el resultado de la investigación se consigna en el estado del arte. Adicionalmente, también se creó un banco de imágenes para análisis el cual se describe más adelante en el presente documento, junto con el método basado en permutaciones.

En la fase de implementación se estudió detalladamente el proceso de descomposición de imágenes en sub matrices y la creación de las permutaciones de la lista principal, así como el proceso para obtener la distribución de probabilidad de las permutaciones en la imagen. Adicionalmente, se desarrollaron las ecuaciones teóricas a utilizar para hallar el valor de entropía y complejidad de acuerdo con lo propuesto en el artículo guía y con las ecuaciones presentadas en el estado del arte. Posteriormente se desarrolló el algoritmo en varias partes: primero se creó el método para calcular la distribución de probabilidad y la graficó, luego el método para calcular entropía y finalmente el método para calcular complejidad. Luego de realizadas las pruebas correspondientes, se unificaron los códigos para obtener un algoritmo que calculara ambas variables recibiendo como parámetro la imagen a analizar. El algoritmo se desarrolló en python, utilizando el computador portátil del ejecutor del proyecto. Por otra parte, las librerías utilizadas fueron elegidas por conveniencia en la simplicidad de las funciones que ejecutan gracias a la revisión de la documentación correspondiente.

Finalmente se aplicó el código a las imágenes del banco de imágenes construido. El análisis se realizó en varias etapas. Inicialmente se evaluó el efecto del formato de la imagen y de las capas de color en el valor de entropía y complejidad calculado. Posteriormente se aplicó

el código a las imágenes de acuerdo con la categoría a la que pertenecían y se obtuvieron resultados en el plano entropía complejidad. Para finalizar, cada una de las imágenes se dividió en 16 trozos de igual tamaño, los cuales fueron analizados y graficados a modo de dispersión en el plano entropía complejidad. Adicionalmente, se contabilizó el tiempo que le toma al código procesar una imagen y devolver el valor de entropía y complejidad. Para finalizar, se realizaron comentarios sobre el código construido y cómo podría mejorarse.

4. Desarrollo de ecuaciones y metodología de permutaciones

A continuación se explicará en que consiste el método para analizar imágenes a partir de permutaciones.

Una imagen en formato digital es una representación bidimensional a partir de la superposición de capas de colores. Las capas de colores se representan con matrices cuyos valores numéricos representan la intensidad del color de cada pixel que conforma la imagen. Con este método se selecciona una de las capas de color y se divide en matrices 2x2. Las matrices se representan como listas que luego se organizan de acuerdo con los índices de los números de la matriz ordenados de menor a mayor. Un ejemplo de este proceso se presenta a continuación:

$$\begin{bmatrix} 23 & 24 & . & . & . & 5 \\ 19 & 22 & . & . & . & 12 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 10 & 22 & . & . & . & 14 \end{bmatrix} \Rightarrow \begin{bmatrix} 23 & 24 \\ 19 & 22 \end{bmatrix} \Rightarrow [23, 24, 19, 22] \Rightarrow [2, 3, 0, 1]$$

Las listas de índices que se obtienen son una permutación de la lista presentada en la ecuación 6. Esta lista tiene 24 posibles permutaciones ya que los valores 0, 1, 2, 3 solamente pueden aparecer una vez en la lista.

$$[0, 1, 2, 3], \quad \text{Permutaciones} = 4 * 3 * 2 * 1 = 24. \quad (6)$$

A continuación, se barreren todas las matrices 2x2 que conforman la capa de color, luego de transformarlas en listas de índices. Con el barrido se encuentra la frecuencia con la que

apareció cada permutación y, a partir de allí, la probabilidad de la permutación. Con la información de las probabilidades para cada permutación es posible calcular la entropía con la ecuación 2. En este caso como se desea que la entropía sea normalizada, se divide en el valor máximo de entropía que se puede tener que, como se indica en las propiedades de la entropía, es el logaritmo natural del número de microestados posibles: 24 permutaciones. De esta forma, la entropía normalizada se calcularía como:

$$H_N(p) = -\frac{1}{\ln(24)} \sum_{i=1}^{24} p_i \ln(p_i). \quad (7)$$

Con respecto a la complejidad, se utiliza la ecuación presentada previamente (ecuación 12) donde p será la distribución de probabilidad obtenida para las 24 configuraciones posibles, y U es la distribución uniforme donde cada permutación tiene una probabilidad de $p(e_k) = 1/24$. En este caso se considera de ambas distribuciones tienen el mismo peso por lo que $\pi(1) = \pi(2) = 1/2$. Entonces los términos para calcular la complejidad son:

$$D(p, U) = \left(H\left(\frac{p+U}{2}\right) - \frac{1}{2}H(p) - \frac{1}{2}H(U) \right). \quad (8)$$

$$H = -\sum_{i=1}^{24} p_i \ln(p_i). \quad (9)$$

$$D^* = -\frac{1}{2} \left(\frac{n+1}{n} \ln(n+1) + \ln(n) - 2\ln(2n) \right). \quad (10)$$

y la ecuación para la complejidad queda finalmente:

$$C(p) = \frac{H_N(p)}{-\frac{1}{2} \left(\frac{n+1}{n} \ln(n+1) + \ln(n) - 2\ln(2n) \right)} \left(H\left(\frac{p+U}{2}\right) - \frac{1}{2}H(p) - \frac{1}{2}H(U) \right), \quad (11)$$

$$C(p) = \frac{H_N(p)}{-\frac{1}{2} \left(\frac{25}{24} \ln(25) + \ln(24) - 2\ln(48) \right)} \left(H\left(\frac{p+U}{2}\right) - \frac{1}{2}H(p) - \frac{1}{2}H(U) \right). \quad (12)$$

5. Descripción banco de imágenes

El banco de imágenes creado para la realización del proyecto puede encontrarse en el repositorio de google drive (Rivera, 2021a). El banco de imágenes cuenta con 5 categorías, 2 pertenientes a ciencia (microscopía y astronomía) y 3 pertenientes a áreas variadas como lo son arte, rostros y la categoría otros que contiene imágenes seleccionadas para probar el código. Se recolectaron un total de 40 imágenes que se distribuyen de la siguiente forma:

- Astronomía: Fueron tomadas 7 imágenes del archivo de cosmología del European Southern Observatory (ESO, 2019).
- Microscopía: Fueron tomadas 10 imágenes provenientes de varias fuentes como la Universidad de Almería (de comunicación, 2020) o el Instituto de ecología (INECOL, 2017).
- Arte: Fueron tomadas 10 imágenes obtenidas de la misma fuente que utilizaron en el artículo guía del presente trabajo (WikiArt, s.f.).
- Rostros: Se recolectaron 10 imágenes de varias fuentes: (Rodriguez, s.f.) (Freepik, 2019) (123RF, 2021) (N.A., 2021) (N.A., 2020). El criterio principal de selección fue que las imágenes tuvieran un fondo lo más uniforme posible. Se seleccionaron rostros de varias etnias.
- Otros: Se seleccionaron 2 ilustraciones y un histograma de autoría propia.

6. Resultados y análisis

Los resultados de la fase de investigación se presentan en las secciones estado del arte, desarrollo de ecuaciones y metodología de permutaciones y Descripción del banco de imágenes.

Con respecto a la fase de implementación, inicialmente se creó un método para calcular la entropía y un método para calcular la complejidad. Posteriormente estos métodos fueron unificados en la versión final del código. Todos los códigos fueron escritos en python utilizando las librerías *matplotlib* y *numpy*. Todos los códigos fueron archivados en un repositorio de github (Rivera, 2021b)

Se creó una función que se encarga de calcular la entropía emulando el procedimiento descrito en la metodología de permutaciones: 1) selecciona una de las capas de color de la imagen a analizar; 2) divide la capa de color en matrices 2x2; 3) convierte las matrices en listas con índices; 4) verifica la frecuencia de cada una de las 24 permutaciones; 5) calcula la probabilidad de cada permutación y con dicha información, 6) calcula la entropía normalizada. La función recibe como parámetro la ubicación de la imagen a procesar, por otra parte, como resultados de la función se obtiene el histograma de frecuencias de las configuraciones y la entropía normalizada. Este código puede encontrarse en el repositorio bajo el nombre *Entropía.ipynb*.

El código para complejidad también fue escrito en python y utiliza las librerías matplotlib y numpy. Este código contaba con dos funciones: la primera calcula la distribución de probabilidad para las 24 configuraciones de acuerdo con los siguientes pasos: 1) selecciona una de las capas de color de la imagen a analizar; 2) divide la capa de color en matrices 2x2; 3) convierte las matrices en listas con índices; 4) verifica la frecuencia de cada una de las 24 permutaciones; y 5) calcula la probabilidad de cada permutación. La distribución de probabilidad es devuelta por dicha función, y es el parámetro que recibe la segunda función, la cual calcula la complejidad de acuerdo con la ecuación 12 y devuelve dicho valor. Este código puede encontrarse en el repositorio bajo el nombre *Complejidad.ipynb*.

Respecto a las pruebas piloto, se realizaron dos con el fin de evaluar si la elección de la capa o el formato afecta o no significativamente el valor de la entropía normalizada y de la complejidad. Los códigos de las pruebas piloto para entropía y para complejidad se encuentran en el repositorio bajo los nombres *Entropía pruebas piloto.ipynb* y *Complejidad pruebas piloto.ipynb*. Con respecto a la capa elegida, se aplicó el método para calcular entropía normalizada y se observó qué tanto difieren entre sí los valores. Dado que todas las imágenes a analizar son en formato jpg, jpeg, y png, cada imagen se transformó a los otros formatos utilizando la librería Pillow. Posteriormente se calculó la entropía en cada caso. Las imágenes utilizadas para realizar estas pruebas se presentan en la figura 1.

Los resultados para la comparación de la entropía en los 3 formatos se presentan en el cuadro 1. Se puede observar que entre el formato jpg y png el valor obtenido para entropía es el mismo hasta en 8 cifras decimales. En el caso del formato jpeg el valor de la entropía se cambia entre la segunda y la tercera cifra decimal. Debido a que prácticamente todas las figuras son en formato jpg o png, la discrepancia que genera el formato jpeg no se considera como importante o como una variable a considerar para el análisis de las imágenes. Por otra parte, existe la posibilidad de solucionar este problema al transformar las figuras en formato

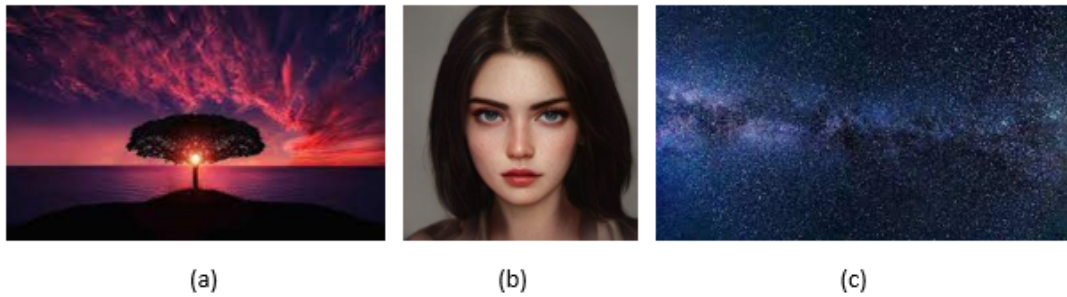


Figura 1: Imágenes utilizadas para pruebas de formato y capas de color. La figura (a) es una expresión artística de un paisaje, la figura (b) es un dibujo de un rostro y la figura (c) es una fotografía del cielo. (a) y (c) vienen de la categoría "otros" en el banco de imágenes, mientras que la imagen (b) proviene de la categoría "rostros".

jpeg a formato png y jpg.

Del resultado para entropía normalizada se puede intuir que el valor de complejidad, que depende de la entropía normalizada y de la entropía de Shannon, será el mismo para las figuras con formato jpg y png, mientras que variará para el formato jpeg. Los resultados para complejidad se presentan en el cuadro 2 y apoyan la suposición anterior. En este caso la complejidad varía entre formatos png, jpg y jpeg a partir de la tercera cifra decimal, mientras que entre png y jpg se mantiene igual para hasta 8 cifras decimales.

Imagen	Entropía jpg	Entropía png	Entropía jpeg
a	0.781	0.781	0.782
b	0.815	0.815	0.831
c	0.998	0.998	0.998

Cuadro 1: Resultados para el ensayo de formatos para las tres imágenes presentadas en la figura 1. La prueba se realizó usando la capa g de todas las figuras.

Imagen	Complejidad jpg	Complejidad png	Complejidad jpeg
a	0.197	0.197	0.196
b	0.002	0.002	0.002
c	0.188	0.188	0.177

Cuadro 2: Resultados para el ensayo de formatos para las tres imágenes presentadas en la figura 1. La prueba se realizó usando la capa g de todas las figuras.

Los resultados para la comparación de la entropía utilizando capas de color se presentan en el cuadro 3. Se puede observar que entre las capas la diferencia en el valor de entropía se presenta a partir de la segunda cifra decimal. Sin embargo la diferencia no se considera significativa. Por el lado de la complejidad, los resultados están en el cuadro 4 y muestran la

misma discrepancia entre capas que en la entropía, pero nuevamente no se considera un cambio muy significativo.

Imagen	Entropía capa g	Entropía capa r	Entropía capa b
a	0.781	0.788	0.791
b	0.815	0.815	0.815
c	0.998	0.998	0.988

Cuadro 3: Resultados para el ensayo de capas de color para las tres imágenes presentadas en la figura 1 en formato png.

Imagen	Complejidad capa g	Complejidad capa r	Complejidad capa b
a	0.182	0.196	0.182
b	0.002	0.002	0.002
c	0.188	0.188	0.186

Cuadro 4: Resultados para el ensayo de capas de color para las tres imágenes presentadas en la figura 1 en formato png.

Previo a realizar el análisis de todas las imágenes se realizaron otros ensayos con el objetivo de verificar que los valores de entropía y complejidad presenten el comportamiento esperado de acuerdo con la distribución de probabilidad. Estos ensayos se realizaron considerando la capa g y utilizando como formato el jpg o el png puesto que generan el mismo valor de entropía para una misma imagen. Los histogramas se crean con los datos de frecuencia para cada permutación de las 24 posibles. Las imágenes con su respectivo histograma se presentan en las figuras 2, 3, 4. Con estas figuras se puede ilustrar que el comportamiento de la entropía se acoge a lo esperado. En el caso de la figura 3, la imagen cuenta con un histograma similar al de una distribución uniforme, y una entropía muy cercana a uno, justo como se esperaba por las propiedades de la entropía de Shannon, mencionada en el estado del arte. Por otra parte, en el caso de la figura 4, el histograma es similar al de una distribución donde un solo microestado tiene toda la probabilidad, por lo que su valor de entropía es muy cercano a cero. Este comportamiento permite señalar que las imágenes con menores microestados posibles (por ejemplo, con una gran región de un solo color) contarán con valores de entropía menores.

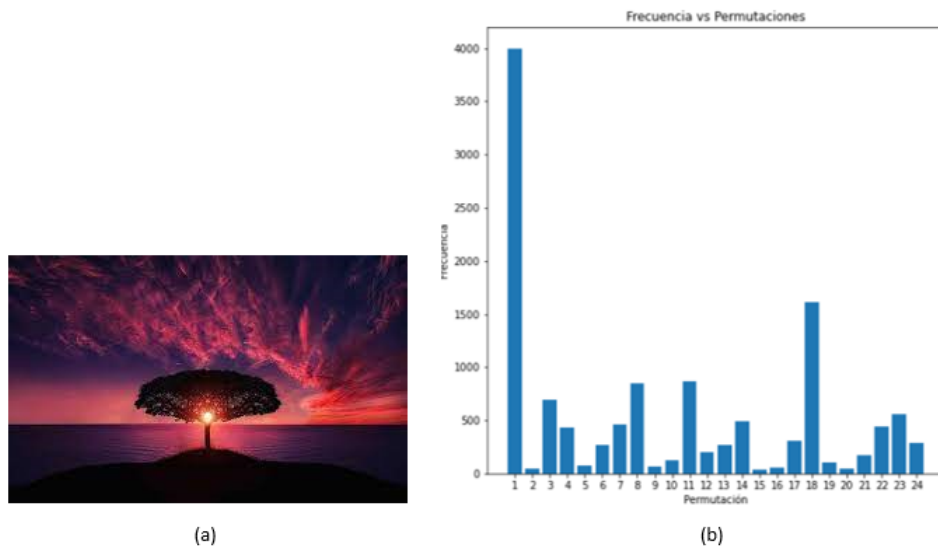


Figura 2: Imagen con el respectivo histograma obtenido con la frecuencia de cada una de las 24 permutaciones. La entropía obtenida es 0.781 y la complejidad es 0.197.

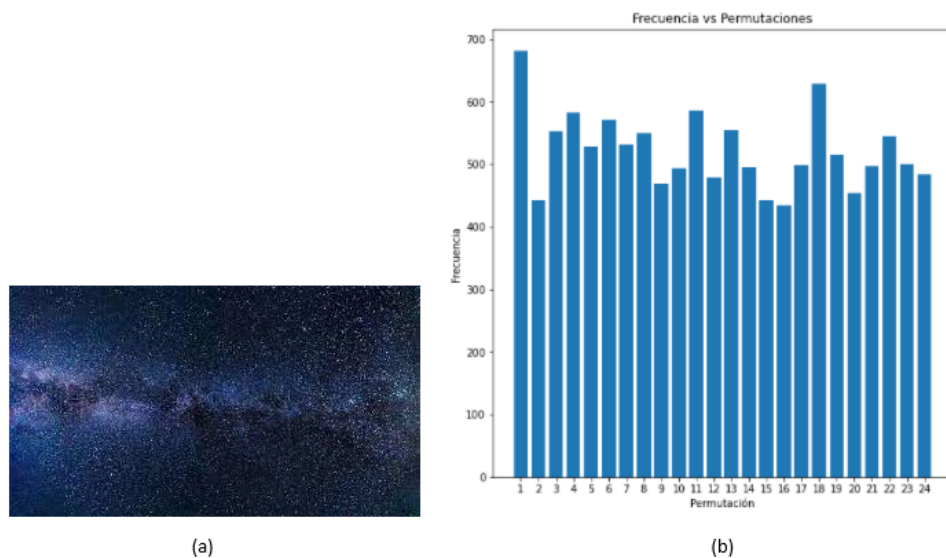


Figura 3: Imagen con el respectivo histograma obtenido con la frecuencia de cada una de las 24 permutaciones. La entropía obtenida es de 0.998 y la complejidad es 0.002.

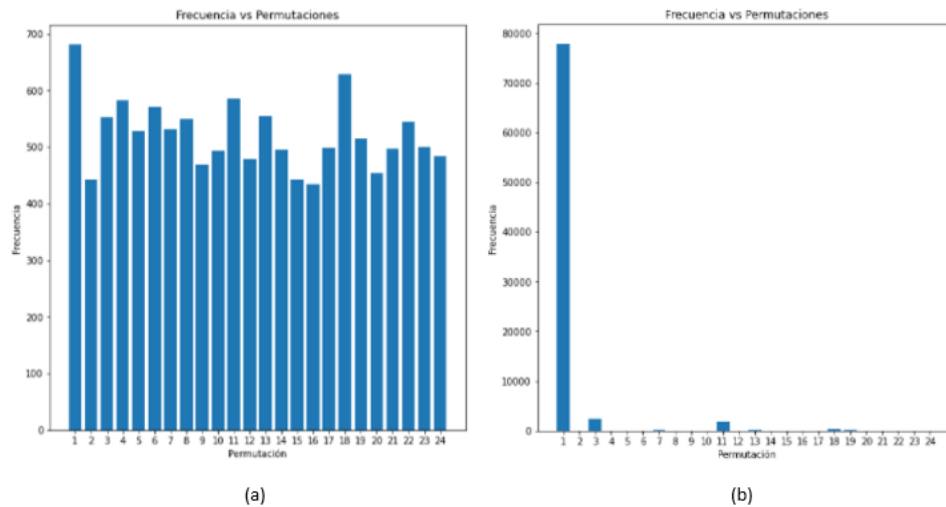


Figura 4: Imagen con el respectivo histograma obtenido con la frecuencia de cada una de las 24 permutaciones. La entropía obtenida es de 0.096 y la complejidad es 0.088

Gracias a las pruebas anteriores se decide trabajar la capa g puesto que las diferencias en valores de entropía y complejidad no se consideran significativas, y manejar todas las imágenes en formato .jpg o .png.

Una vez superada la etapa de implementación viene la etapa de aplicación. Para esta etapa se modificaron los códigos de entropía y complejidad. Estos se unificaron en uno solo creando dos funciones, la primera obtiene la distribución de probabilidad para las permutaciones, y la segunda calcula la entropía normalizada y la complejidad. El código se puede encontrar en el repositorio bajo el nombre *Entropía-Complejidad.ipynb* (Rivera, 2021b). Este código cumple la función de devolver una lista con el valor de la entropía y de complejidad contenido en ella. Posteriormente se aplicó el código a las imágenes recopiladas, agrupando los resultados de acuerdo con las categorías, es decir, se generó un plano de entropía vs complejidad para cada una de las categorías del banco de imágenes. Se puede observar en la figura 5 el resultado de aplicar el código desarrollado a todas las imágenes. Se observa una tendencia general a agruparse en valores de entropía de medios a elevados y de complejidad bajos. Obtener valores de entropía elevados significa que en la mayoría de las imágenes hay una tendencia a encontrar una distribución mas o menos cercana a la uniforme para las permutaciones de las matrices. Por otro lado, valores de complejidad bajos muestran que, en la mayoría de las imágenes debe existir, bien sea, un patrón o un comportamiento ruidoso. También se observa una tendencia generalizada a disminuir el valor de la complejidad conforme el valor de la entropía aumenta. Esto tiene sentido en todos los casos porque, de acuerdo con la

teoría, mayor entropía significa distribución de probabilidad cercana a la uniforme, y una distribución uniforme puede asociarse aun comportamiento aleatorio o ruidoso.

El resultado por categoría se presenta en la figura 6. Para el caso de las obras de arte, la entropía oscila entre 0.74 y 1. Para astronomía la mayor parte de los datos oscila entre 0.8 y 1 para entropía y entre 0 y 0.2 para complejidad, sin embargo una figura genera un valor de entropía de 0.53 y un valor de complejidad de 0.257. Para las imágenes de microscopía la entropía oscila entre 0.75 y 1 y la complejidad entre 0 y 0.2. Finalmente, para la categoría rostros la entropía oscila entre 0.65 y 1, y la complejidad entre 0 y 0.2.

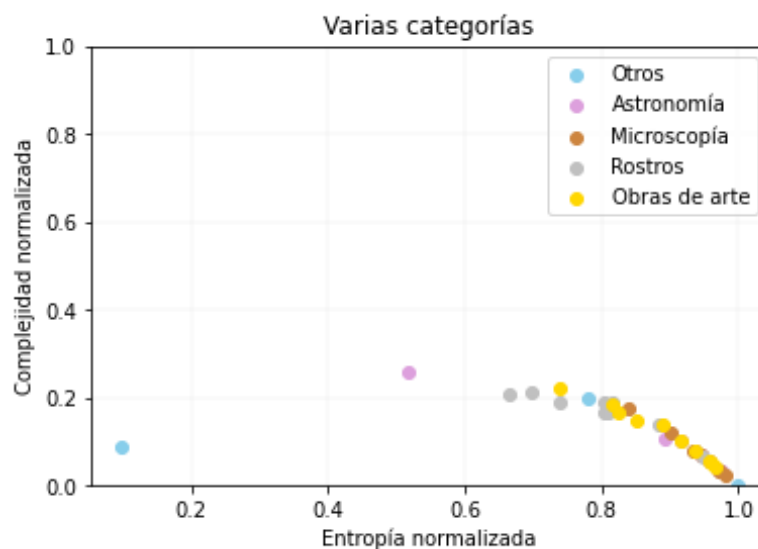


Figura 5: Plano Entropía-Complejidad con los resultados de todas las categorías.

En el caso de astronomía, se esperaba la tendencia a valores de entropía elevados porque en las imágenes son pocas las zonas "uniformes" en color, que puedan ocasionar que una permutación se lleve toda la probabilidad; con respecto a complejidad, también se esperaba la tendencia a valores bajos puesto que las imágenes representan redes, lo que puede entenderse como un patrón. En el caso de microscopía también se esperaban valores elevados de entropía puesto que las imágenes contienen, en algunos casos muchas formas, lo que hace que la descomposición en matrices de permutaciones más variadas principalmente en los límites de las formas; en los otros casos las imágenes presentan pocas zonas uniformes en color, por lo que también se obtienen frecuencias de permutaciones distribuidas; con respecto a complejidad, obtener valores bajos también era esperable puesto que las imágenes son observaciones de organismos o estructuras, por lo que presentan algún patrón. Con respecto a los rostros, se esperaban valores intermedios de entropía puesto que se seleccionaron

imágenes con fondos de un único color, es decir, imágenes con una gran zona uniforme que haría concentrar la frecuencia en pocas permutaciones, generando así un valor de entropía intermedio; el hecho de que la entropía sea elevada se atribuye entonces a las facciones del rostro y a los bordes entre rostro y fondo que hacen la distribución de frecuencias más cercana a la uniforme. Por el lado de la complejidad para rostros, obtener valores bajos también fue sorprendente puesto que, aunque los rostros presentan cierta simetría, no puede establecerse un patrón, como lo sería una red, o considerarse una imagen ruidosa puesto que sí cuenta con una figura. Sin embargo, el hecho de que la complejidad sea baja responde a que la entropía es alta lo que va de acuerdo con la teoría previamente estudiada. De haber obtenido los valores intermedios esperados, esto también habría estado de acuerdo con la teoría. Finalmente, las obras de arte reprodujeron el patrón mostrado en el artículo: valores de entropía elevados generan valores de complejidad bajos.

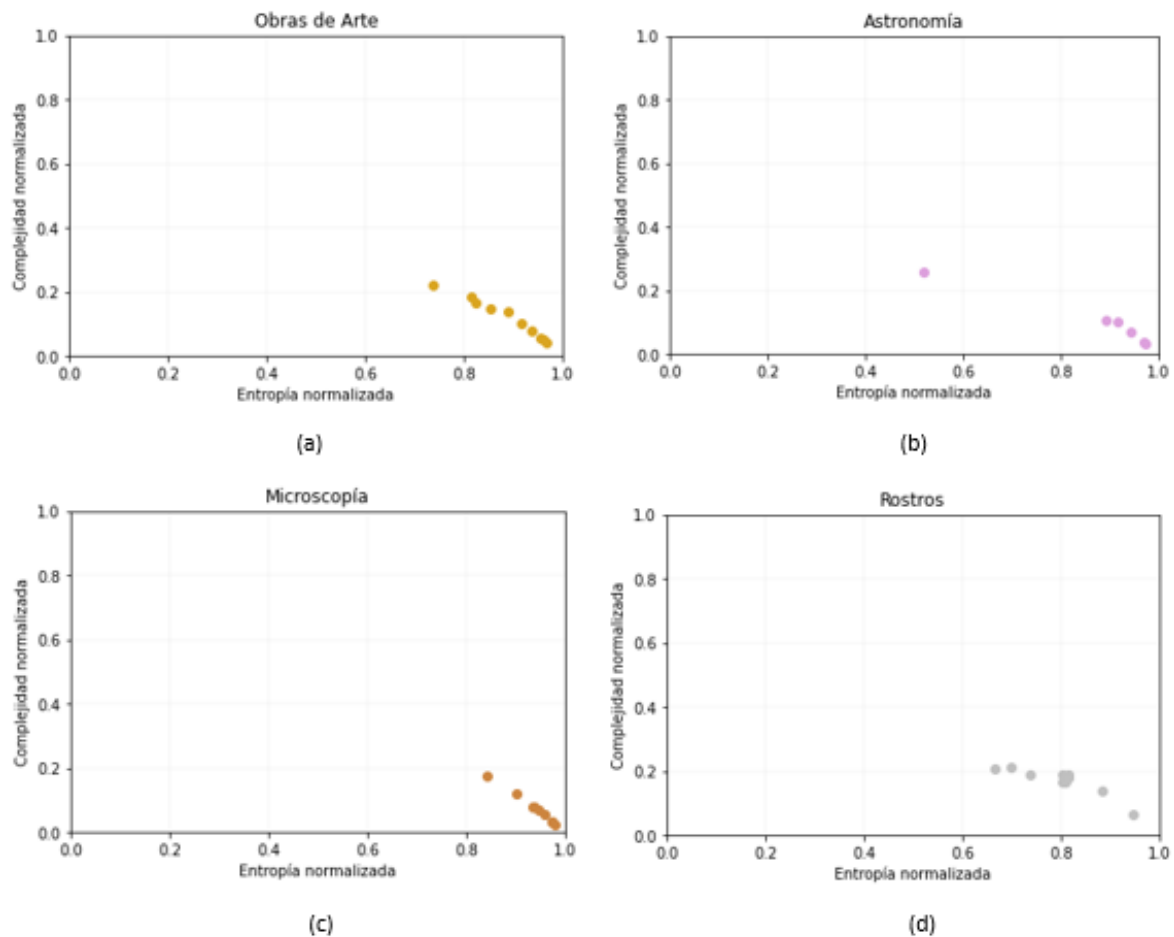


Figura 6: Plano Entropía-Complejidad con los resultados para cada categoría. (a) Son los resultados para obras de arte, (b) son los resultados para astronomía, (c) son los resultados para microscopía y (d) son los resultados para rostros.

Como análisis adicional se descompusieron cada una de las imágenes en 16 subimágenes a las que se les aplicó nuevamente el código (las descomposiciones también se encuentran en el banco de imágenes (Rivera, 2021a)), esto con el objetivo de verificar si las distintas zonas de la imagen se ubicaban cerca al valor de entropía complejidad de la imagen final total. Lo que se obtuvo entonces fue una dispersión para cada categoría, conformada por las imágenes descompuestas. Los resultados para astronomía se presentan en la figura 7, para microscopía en la figura 8, para rostros en la figura 9 y para arte en la figura 10.

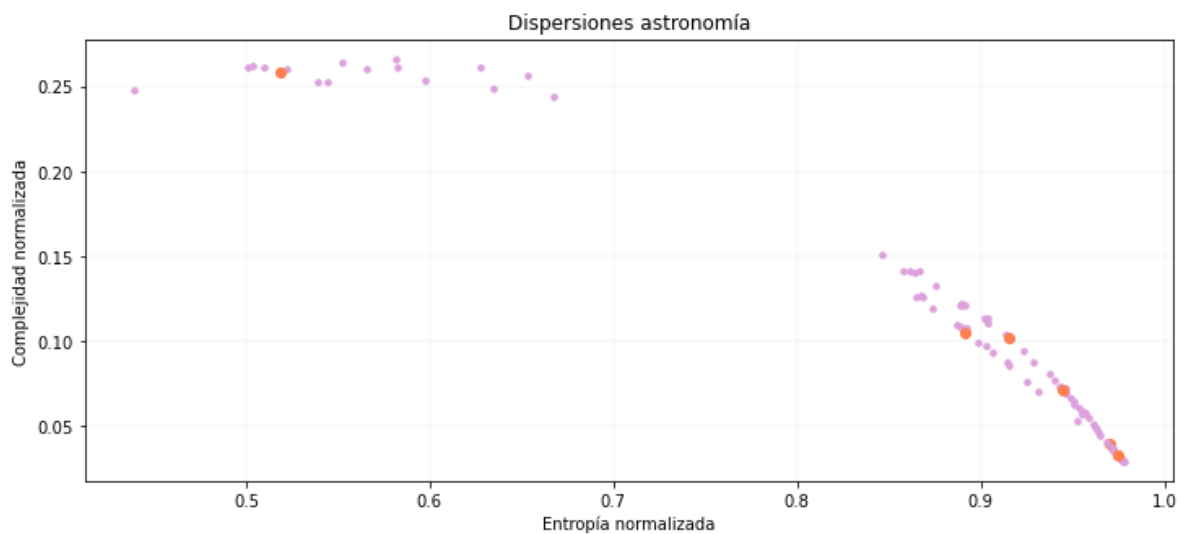


Figura 7: Plano Entropía-Complejidad con la dispersión generada por la división de las imágenes en 16 subimágenes para la categoría astronomía. En rosa las imágenes completas, en morado la dispersión producida por las imágenes.

Para astronomía y microscopía las nubes de puntos responden a la tendencia general de la categoría, valores elevados de entropía y valores bajos de complejidad, aunque las dispersiones superan los valores límites encontrados al trabajar con las imágenes completas. Esto puede deberse a que las subimágenes pueden encerrar zonas dónde no existan patrones claros (alterando el valor de la complejidad), o dónde existan zonas uniformes en color (alterando el valor de la entropía). En el caso de astronomía, se marcan dos zonas que concentran nubes de puntos: al rededor de la mayoría de los puntos y al rededor del punto que cuenta con un valor de entropía menor y un valor de complejidad mayor que las demás. Esto hace que la tendencia de la gráfica no se vea lineal sino parabólica, aunque para esto no se tiene una explicación. Por otra parte, en el caso de microscopía, la tendencia incluyendo la nube de puntos sigue siendo lineal por lo que para mayores valores de entropía se tienen menores valores de complejidad.

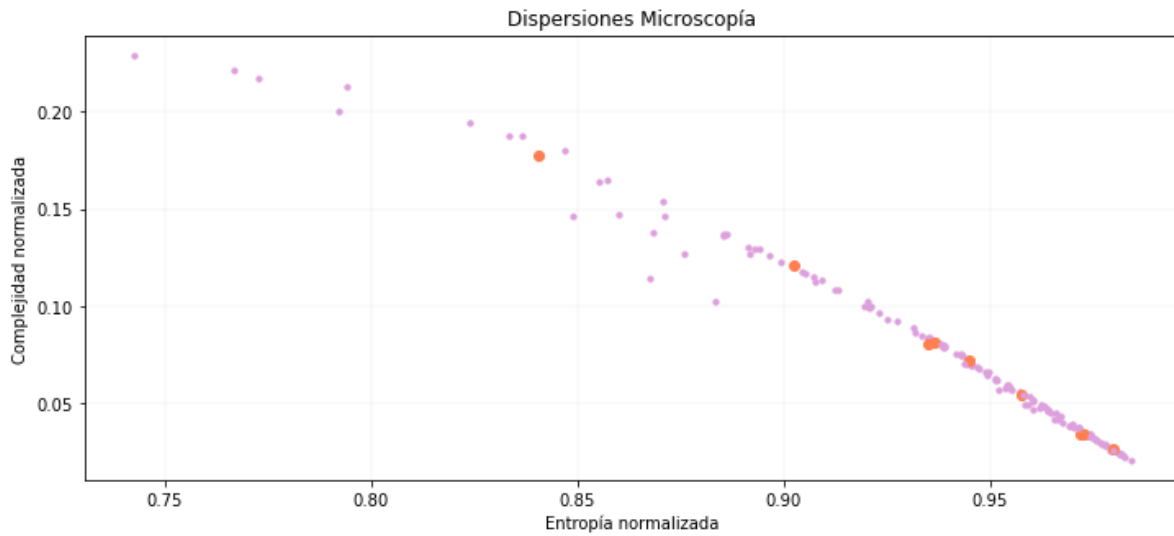


Figura 8: Plano Entropía-Complejidad con la dispersión generada por la división de las imágenes en 16 subimágenes para la categoría microscopía. En rosa las imágenes completas, en morado la dispersión producida por las imágenes.

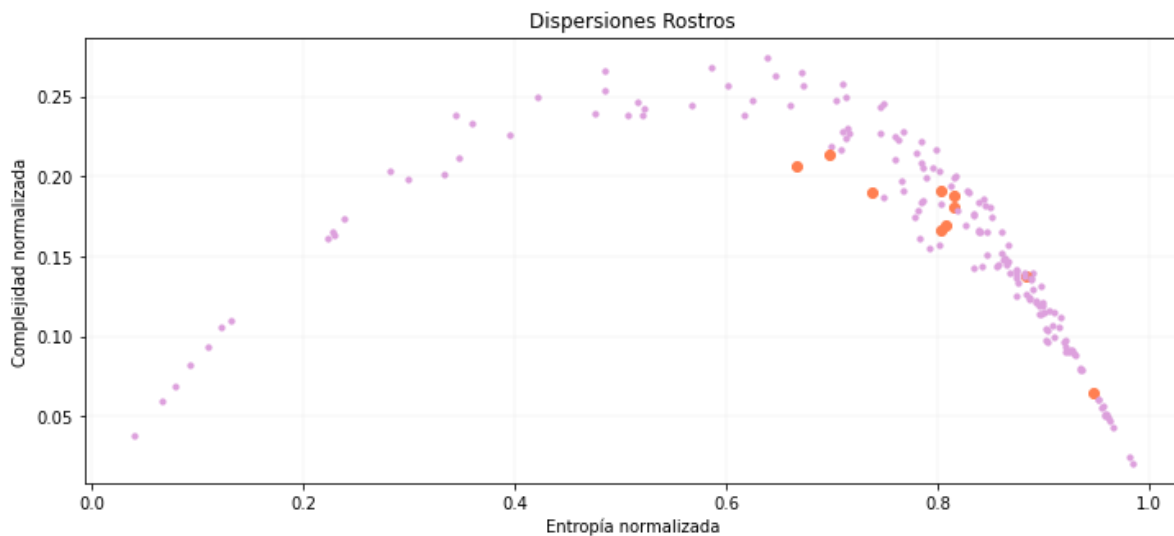


Figura 9: Plano Entropía-Complejidad con la dispersión generada por la división de las imágenes en 16 subimágenes para la categoría rostros. En rosa las imágenes completas, en morado la dispersión producida por las imágenes.

Para la categoría rostros, se observa una tendencia parabólica que se asocia con las características de las figuras seleccionadas: hay zonas dónde la imagen es uniforme en color, lo que va a generar distribuciones dónde la mayor parte de la probabilidad la lleva un microestado, y por lo tanto se producen valores de entropía cercanos a cero; mientras que

también hay otras zonas dónde las imágenes cuentan con variaciones bruscas en tonalidad que generan distribuciones más uniformes y valores elevados de entropía. Con respecto a la complejidad, a pesar de que las zonas uniformes en color tengan valores bajos de entropía, esto no repercute en la complejidad. En esta categoría las dispersiones no siguen la tendencia de las figuras completas: Valores altos de entropía producen valores bajos de complejidad.

Finalmente, para las obras de arte las dispersiones de las imágenes principales mantienen la tendencia de altos valores de entropía producen bajos valores de complejidad. Otro aspecto relevante es que entre más disminuye el valor de entropía, mayor dispersión hay en el valor de complejidad, aunque no se asocia este comportamiento con algo en particular.

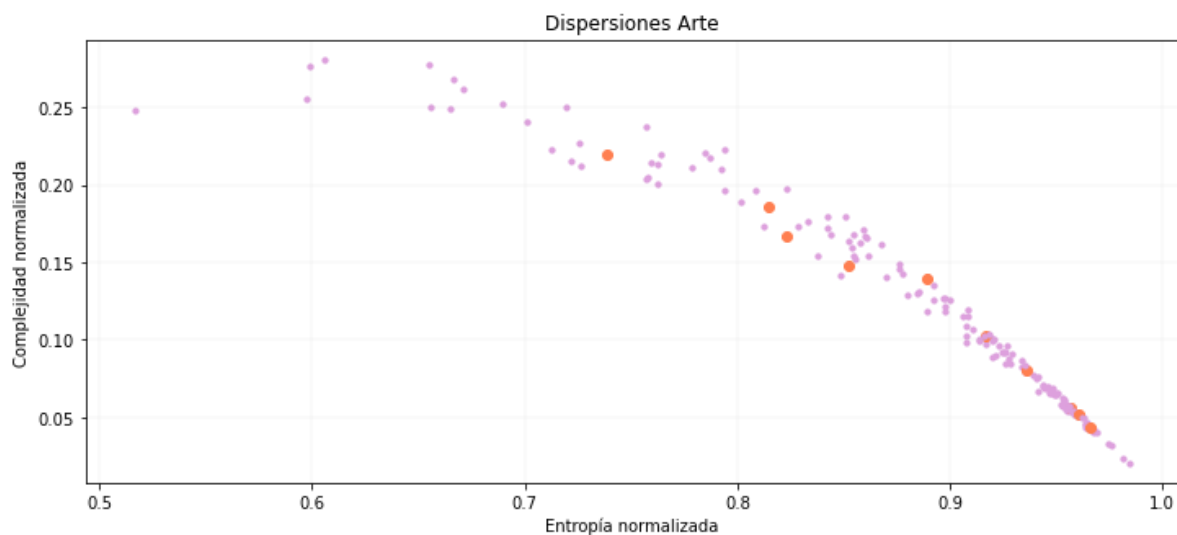


Figura 10: Plano Entropía-Complejidad con la dispersión generada por la división de las imágenes en 16 subimágenes para la categoría obras de arte. En rosa las imágenes completas, en morado la dispersión producida por las imágenes.

Con respecto a la evaluación del código se seleccionó la librería `matplotlib` sobre la librería `pillow` para el procesamiento de imágenes y generación de gráficas porque permite obtener las capas de color de forma más directa puesto que no requiere realizar ninguna operación sobre la figura, mientras que la librería `pillow` si lo requiere (vease código probando las dos técnicas en el archivo *Borrador 2.ipynb*). Esta prueba se realiza para la imagen más grande que se procesó, la imagen 4 de astronomía y se realiza solamente con el valor de entropía puesto que la complejidad es una función de la entropía. Esta selección genera una diferencia de tiempo de 34s (con `pillow` tarda 3:08 min, con `matplotlib` tarda 2:34 min).

Se considera que el código puede ser mejorado solamente en el proceso de calcular la

distribución de probabilidad, puesto que la siguiente función solamente realiza operaciones algebraicas con los valores encontrados. Podría tratar de realizarse una sola función, pero se descarta esta posibilidad porque al contar con dos funciones se tiene la posibilidad de graficar fácilmente el histograma de frecuencias, y de esta manera verificar que los valores obtenidos tengan sentido. La forma de mejorar el cálculo de la distribución de probabilidad sería utilizando de forma más eficientes los diccionarios, que son la estructura utilizada para almacenar la información de la frecuencia de las permutaciones (véase el repositorio (Rivera, 2021b)).

7. Conclusiones

Con este proyecto se aprendió un nuevo método para calcular entropía utilizando la definición dada por Shannon, y se implementó una definición de complejidad basada en el cálculo de la entropía. Se verificó que el método basado en permutaciones se puede utilizar satisfactoriamente sobre imágenes bidimensionales, tanto provenientes de ciencia como provenientes de otros ámbitos. Adicionalmente, se encontró que la entropía y complejidad definidas teniendo en cuenta el método de permutaciones pueden ser asociadas con características de una imagen: la entropía puede asociarse con la distribución de probabilidad, tomando valores máximos o mínimos dependiendo de si se tiene una distribución uniforme o una distribución donde un solo microestado cuenta con toda la probabilidad; y la complejidad puede asociarse con la existencia de patrones o de comportamientos ruidos en las figuras analizadas. Adicionalmente se pudo observar que la influencia de la capa de color sobre los valores de entropía y complejidad no es muy significativa y que por lo tanto se puede trabajar con cualquiera de las capas que conforman la imagen. También se encontró que existen dos formatos que, bajo el método desarrollado en el presente trabajo, producen los mismos valores de entropía y complejidad: .jpg y .png.

Fue posible construir las curvas de entropía y complejidad y se encontró que la tendencia general, independiente de la categoría, es obtener valores de complejidad pequeños, cuando se tienen valores de entropía elevados. Lo anterior está de acuerdo con el comportamiento esperado al observar los resultados del artículo guía. Del comportamiento de las gráficas se pudo corroborar también el comportamiento esperado por la teoría: si la entropía es elevada, significa que se tiene una distribución uniforme, normalmente observable como ruido, la cual generará valores de complejidad bajos porque, si bien no se tiene un patrón, se tiene un comportamiento ruidoso. Por otra parte, se pudo observar gracias a las pruebas piloto y a las

dispersiones, que las imágenes con zonas uniformes de color van a tener valores de entropía intermedios o menores debido al efecto que estas zonas tienen sobre la distribución de color.

Finalmente, con respecto a la implementación computacional, se encontró que las librerías trabajadas pueden tener implicaciones en la calidad del código, principalmente en términos de tiempo. En este sentido, se determinó que matplotlib permite realizar un mejor procesamiento de imágenes que la librería pillow. Por otra parte, se considera que la mejor forma de optimizar el código desarrollado es modificando el método para obtener la distribución de probabilidad de las permutaciones aquí estudiadas, mejorando principalmente el uso de los diccionarios.

8. Bibliografía anotada

- History of art paintings through the lens of entropy and complexity: Se trata de el artículo principal del proyecto. En el se basan los métodos para estudiar entropía y complejidad.
- A mathematical theory of communication: Artículo publicado por Shannon dónde se relaciona la entropía con información y se sientan las bases para la creación de la teoría de la información. Fundamental para conocer propiedades, relación con información y contexto histórico en el que surge. Útil para definir antecedentes en análisis de señales y su comprensión a nivel de procesamiento de información. Es muy antiguo pero su relevancia sigue siendo mucha sobretodo para un trabajo en el que se va a utilizar como principal característica de análisis.
- La gestión de la información en un enfoque a partir de la entropía: El artículo publicado en julio de 2020 cuenta con 1 citación a la fecha, está publicado en la revista de libre acceso "Sinapsis" en Ecuador, revista poco conocida a nivel mundial. Los autores son profesores universitarios en Chile, Ecuador y Cuba. Dos son de nivel PhD y uno es de nivel Maestría. No se identifican sesgos con respecto al estudio realizado. Su redacción, formato y ordenamiento son de calidad. Es útil porque se trata de una revisión teórica del concepto de entropía visto desde varios campos: física, matemática y teoría de la información. Además revisa las ventajas y limitaciones que tiene la entropía para análisis de datos.
- Statistical and Thermal Physics: Útil para la comprensión del concepto de entropía a

nivel físico y sus diferentes interpretaciones. También es útil para comprender propiedades básicas de la entropía y cómo se utiliza para inferir probabilidades. Adicionalmente, ayuda a señalar la relevancia de la entropía en el análisis de sistemas físicos por medio del uso de esta cantidad en la física estadística. Se trata de un libro, utilizado típicamente como libro base para cursos como física estadística o como apoyo en cursos de estado sólido. Es medianamente reciente, de 2010.

- **Análisis de Métodos de Medición de Complejidad de Imagen-Image Complexity Measurement Methods: A Survey:** Se trata de una revisión de las medidas de complejidad en imágenes y la interpretación. Es un artículo bastante reciente (2018) publicado en revista relacionada con informática, computación y similares.
- **Entropía de permutación para la caracterización de texturas en imágenes digitales:** Sirve como antecedente para una aplicación del método de permutaciones. Es un artículo relativamente reciente. Utiliza el concepto de permutaciones para calcular entropía con una "celda" base al igual que se propone en el artículo principal.
- **Entropy and information theory: Uses and misuses:** El artículo fue publicado en el 2019 y cuenta con 11 citaciones a la fecha, se trata de un artículo de acceso libre. La autora fue profesora universitaria de la universidad Hebrea de Jerusalen y ha publicado un libro relacionado con entropía y teoría de la información que ha contado con más de 400 citaciones. Es relevante porque estudia la interpretación que se le da a la entropía en diferentes áreas como la física o la teoría de la información. Por otra parte, presenta ejemplos que permiten comprender qué se puede entender e interpretar a partir de la entropía y qué no.
- **Information theory, inference and learning algorithms:** Es importante para la comprensión del concepto de entropía, su descripción matemática y sus propiedades. También es fundamental para la comprensión de la relación entre entropía e información. Contiene algunos ejemplos de la aplicación de entropía en relación con teoría de información por lo que contribuye al estado del arte.
- **Teoría de la información de Claude E. Shannon:** Se trata de una revisión detallada del artículo original de Shannon, así como un recorrido histórico que llevó a los desarrollos de Shannon y de algunas aplicaciones previas en procesamiento de información. También menciona un ejemplo de la información mutua y cómo esta podría relacionarse con complejidad.

- Breve historia del descubrimiento de la estructura del ADN: Revisión histórica de los sucesos que llevaron al descubrimiento de la estructura del ADN con patrones de difracción de rayos x y cómo el análisis de imagen a partir de intensidad fue fundamental. Contribuye al estado del arte para análisis con base en estudios de intensidad.
- Espectroscopia IR de estrellas B con líneas en emisión: Contribuye al estado del arte como un ejemplo de la utilización del método clásico de análisis basado en intensidades. Es una tesis para obtener un grado de doctor en astronomía y es relativamente reciente, de 2010. También es relevante porque muestra que aunque las técnicas basadas en intensidad sean antiguas, siguen siendo fundamentales en áreas como la astronomía.
- Generalización de la divergencia de Jensen Shannon a estadística no extensiva para el análisis de secuencias: En su introducción contiene información teórica fundamental para la que interpretación se les da a las divergencias. Por otra parte, contribuye a la comprensión de las ecuaciones presentadas en el artículo principal puesto que las analiza término a término. Por otra parte, menciona algunas cantidades que pueden ser definidas en términos de divergencias. Es un artículo relativamente reciente, de 2013.
- Divergence measures based on the Shannon entropy: Contribuye a una comprensión del concepto de divergencia y cómo esta se define a partir de la entropía de Shannon. Brinda una relación evidente entre entropía y divergencia por lo que muestra la relación directa entre la complejidad (en términos de la divergencia) y la entropía. Señala algunas propiedades de la divergencia. Se trata de un artículo antiguo, cercano a la publicación original de Shannon.
- Aplicaciones de las medidas de complejidad estadística: Contribuye con la comprensión del concepto de complejidad estadística y como se ha utilizado. Por lo tanto, contribuye tanto a la teoría como al estado del arte. Ayuda a comprender mejor las ecuaciones desarrolladas y las razones por la que tiene sentido plantearse una expresión para complejidad que dependa de entropía, y por lo tanto, de información.
- Fundamentals of statistical and thermal physics: Contribuye a la comprensión del concepto de entropía. Fundamental por la definición de propiedades y por la evidencia de relevancia con aplicaciones a la física estadística. Permite comprender por qué la entropía ha sido una cantidad tan utilizada para el estudio de sistemas.
- Identificación de metales y no metales en las atmósferas de las estrellas: Un proyecto

de colaboración docente del Instituto de Astronomía y del CCH Naucalpan para el curso de Química I: Contribuye al estado del arte ya que es un ejemplo de aplicación de análisis de intensidad.

Referencias

- 123RF. (2021). https://es.123rf.com/photo_5067834_beb%C3%A9-hermoso-rostro-en-el-fondo-blanco.html.
- Antonelli, A. P. (2017). *Entropía de permutación para la caracterización de texturas en imágenes digitales* (Tesis Doctoral no publicada). Universidad Nacional de Mar del Plata. Facultad de Ingeniería. Argentina.
- Bussandri, D., Linck, L. G., Ré, M., y Lamberti, P. (2014). Generalización de la divergencia de jensen shannon a estadística no extensiva para el análisis de secuencias. En *Anales afa* (Vol. 24).
- de comunicación, G. (2020). *Nuevo microscopio electrónico de barrido de alta resolución de emisión de campo en la ual*. <https://news.ual.es/ciencia/nuevo-microscopio-electronico-de-barrido-de-alta-resolucion-de-emision-de-campo-en-la-ual/>. Universidad de Almería.
- ESO. (2019). *Image archive: Cosmology*. <https://www.eso.org/public/images/archive/category/cosmology/list/2/>. European Southern Observatory.
- Espín, J. C., Toledo, H. H., Vázquez, L. M., Hernández, J. M., y Cielo, A. P. (2015). Identificación de metales y no metales en las atmósferas de las estrellas: Un proyecto de colaboración docente del instituto de astronomía y del cch naucalpan para el curso de química i. *Latin-American Journal of Physics Education*, 9(1), 6.
- Fierro, A. (2001). Breve historia del descubrimiento de la estructura del adn. *Rev Méd Clínica Las Condes*, 20, 71–75.
- Freepik. (2019). *Retratos primer plano*. https://www.freepik.es/fotos-premium/retrato-hombre-albino-estudio_8338342.htm.
- Granada, A. (2010). *Espectroscopia ir de estrellas b con líneas en emisión* (Tesis Doctoral no publicada). Universidad Nacional de La Plata.
- Herrera, L. M., Murguía, M. I. C., y Quintana, J. A. R. (2018). Análisis de métodos de medición de complejidad de imagen-image complexity measurement methods: A survey. *ReCIBE, Revista electrónica de Computación, Informática, Biomédica y Electrónica*, 7(2), 17–46.
- Holík, F. (2016). Teoría de la información de claud e. shannon. *Diccionario Interdisciplinar Austral*, 1, 1–17.
- INECOL. (2017). *Microscopía electrónica de barrido y microanálisis de elementos del clúster científico y tecnológico*. <https://www.inecol.mx/inecol/index.php/es/ct-menu-item-25/ct-menu-item-27/17-ciencia-hoy/723-microscopia-electronica-de-barrido-y-microanalisis-de-elementos-del-cluster>

- cientifico-y-tecnologico-biomimic. Instituto de ecología, A.C.
- Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1), 145–151.
- MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.
- Morel, C., y Humeau-Heurtier, A. (2021). Multiscale permutation entropy for two-dimensional patterns. *Pattern Recognition Letters*, 150, 139–146.
- N.A. (2020). *Blanqueamiento de rostro*. <https://wapa.pe/moda-y-belleza/1549241-truco-chinas-rostro-blanco-piel-mascarilla-belleza-rostro>. WAPA.
- N.A. (2021). *Cantata*. <https://co.pinterest.com/pin/752804893948846453/>. Pinterest.
- Reif, F. (1998). *Fundamentals of statistical and thermal physics*. American Association of Physics Teachers.
- Rivera, K. (2021a). *Imágenes proyecto computacional*. https://drive.google.com/drive/folders/1bSAK3gGUHBPYJrc_jNTn4FPL1TgbK0aU?usp=sharing. Google.
- Rivera, K. (2021b). *Proyecto teórico*. <https://github.com/karolyrr/Proyecto-te-rico>. GitHub.
- Rodriguez, G. (s.f.). *Modelo y belleza peculiar*. <https://i.imgur.com/XpgLq76.png>. recreo-viral.
- San Martín, V. M. (2013). Aplicaciones de las medidas de complejidad estadística.
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3), 379–423.
- Sigaki, H. Y., Perc, M., y Ribeiro, H. V. (2018). History of art paintings through the lens of entropy and complexity. *Proceedings of the National Academy of Sciences*, 115(37), E8585–E8594.
- WikiArt. (s.f.). <https://www.wikiart.org/>. Visual Art Encyclopedia.