VOLUMEN 37
XXX Verano De la Ciencia
ISSN 2395-9797
www.jovenesenlaciencia.ugto.mx

Evaluación de modelos de difusión para la estimación de densidad de multitudes a partir de videos.

Evaluation of Diffusion Models for Crowd Density Estimation from Videos.

C. Álvarez Arredondo[1], J. B. Hayet[2]

[1]Estudiante de Licenciatura en Física. División de Ciencias e Ingenierías. Campus León. Universidad de Guanajuato

c.alvarezarredondo@ugto.mx

[2] Centro de Investigación en Matemáticas (CIMAT) Guanajuato, Gto, México.

jbhayet@cimat.mx

Resumen:

El **análisis de multitudes** ha sido de gran relevancia en los últimos años, con aplicaciones en cuestiones de seguridad, de monitoreo urbano, gestión de eventos masivos o análisis demográfico. La modelación del movimiento de multitudes y el desarrollo de técnicas de predicción o de simulación basados en aprendizaje máquina, a partir de datos, permiten aportar estrategias para resolver este problema, y por ende mejorar la seguridad y la prevención de riesgos en los contextos antes dichos. Al día de hoy, el **aprendizaje maquina** es la herramienta principal para poder desarrollar este objetivo, y, entre todas las técnicas modernas de aprendizaje máquina, los **procesos de difusión** ofrecen una alternativa poderosa en la estimación de densidad a partir de imágenes de multitudes. Estos modelos están inspirados en procesos estocásticos de tipo Márkov que aprenden a generar datos complejos, mediante la perturbación por ruido de los datos de entrenamiento y la optimización de los parámetros de una red neuronal para estimar el ruido aplicado. Sabiendo cómo estimar el ruido aplicado en un cierto paso, estos modelos generativos pueden sintetizar datos plausibles a partir de puro ruido. En nuestro caso, los modelos de difusión aprenden a generar una representación visual de la densidad asociada a una imagen de una multitud, a partir de ruido puro y de manera condicionada a la imagen observada. Posteriormente, se analiza para estimar el número de personas presentes en la escena. En este trabajo, se explora la aplicación de estos modelos con datos del **Festival de Love Parade** del 2010 en Duisburgo (Alemania), en el cual ocurrió una estampida, dejando aproximadamente 51 muertos. Sucesos como estos son la motivación para realizar este tipo análisis para prevenir accidentes fatales.

Palabras clave: Aprendizaje máquina; Modelos de difusión; Denoising process; Mapas de densidad; Análisis de imágenes de multitudes.

Introducción:

Con una triste regularidad, noticias trágicas en todo el mundo nos recuerdan que las congregaciones masivas de personas están relacionadas a grandes riesgos de accidentes y pérdidas de vidas humanas, por movimientos masivos pudiendo resultar en estampidas. En el estado de Guanajuato, eventos como la feria de León o el Día de las flores en la capital, muy apreciados por la población local y cada vez más por poblaciones foráneas, son el teatro de congregaciones de miles de personas, donde se alcanzan altos niveles de densidad.

En este contexto, la modelación del movimiento de multitudes y el desarrollo de herramientas de predicción basadas en estos modelos, son cruciales: Por un lado, permiten aportar elementos de evaluación para un diseño adecuado de los entornos arquitectónicos, para la elección de estrategias de escape en caso de evacuación, es decir para todo trabajo de preparación de un evento masivo y prevención de los riesgos asociados. Estamos convencidos de que los modelos desarrollados podrían llevarnos en un futuro cercano a la implementación de sistemas de monitoreo y de anticipación (en vivo), capaces de emitir alarmas frente a

situaciones susceptibles de resultar en posibles compresiones dentro de la multitud observada. El trabajo realizado en este proyecto propone modestamente hacer un paso en esta dirección.

La mayoría de los trabajos pasados han hecho uso de los llamados modelos físicos, es decir modelos donde leyes de la física se aplican para la evolución de un conjunto de partículas (las personas de la multitud), capturando fenómenos de tipo evitamiento o seguimiento; ahora, desde hace algunos años, la llegada de nuevos métodos de aprendizaje máquina, con aprendizaje profundo, y la disponibilidad de bases de datos nuevas ha abierto la puerta a la resolución del problema a partir de datos de entrenamiento. En este trabajo, nos enfocamos en una parte del problema: La implementación de métodos permitiendo generar **mapas de densidad** (idealmente, con datos de densidad expresados en personas/metro cuadrado) a partir de simples imágenes de multitudes.

Como lo describiremos en este documento, se parte de una imagen como entrada y se emplea un **modelo de difusión entrenado para poder generar una salida visual** (imagen de puntos o regiones) donde cada punto o elemento representa a una persona estimada de la imagen inicial. Estos modelos generativos exhiben cierta robustez en escenas complejas y logran una interpretación visual sencilla para el conteo. La meta es poder construir una representación del estado de la multitud como ilustrado en la Figura 1, es decir una representación de tipo **rejilla métrica**, con valores de densidad y de velocidad local asociados a cada elemento de esta rejilla. Esta representación, en un trabajo futuro, nos servirá de base para aprender a predecir el movimiento de la multitud.

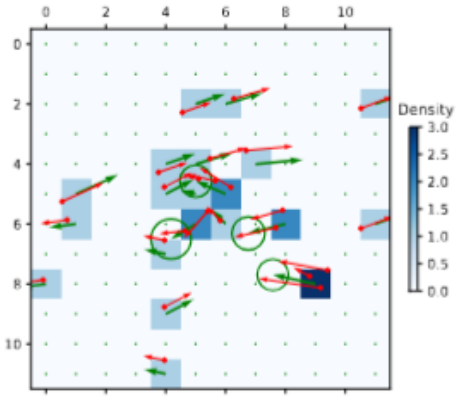


Figura 1. Representación de una multitud por mapa de densidad.

El trabajo propuesto se enfocó en dos aspectos:

- la **estimación de densidad** a partir de un cuadro del vídeo;
- la **compensación de la perspectiva** para poder generar mapas métricos.

Modelos generativos de difusión:

A continuación, damos unos elementos explicativos sobre lo que constituye la esencia del modelo usado, el modelo de difusión.

Para poder utilizar adecuadamente un modelo de difusión, es necesario preprocesar los datos (en nuestro caso, las imágenes) antes de aplicar el modelo. Cada imagen es redimensionada a una representación estándar de 256x256 pixeles y normalizada para poder adecuarse al dominio del modelo. En este trabajo, se utiliza un modelo DDPM (*Denoising Diffusion Probabilistic Models*); esto quiere decir que dado un dato de entrenamiento x_0 se define un proceso de Márkov el cual agrega ruido gaussiano de manera progresiva:

$$q(x_t|x_{t-1}) = N\left(x_t; \sqrt{(1-\beta)}x_{t-1}, \beta_t I\right).$$

Esta expresión nos da la distribución del dato "más ruidoso", dado el dato del paso anterior, para un paso del proceso de **difusión hacia adelante**. Generalmente, este proceso se conoce como "*forward process*" donde $N(\dots)$ denota una distribución normal, β_t es el parámetro de varianza del ruido agregado, el cual debe ser pequeño y positivo ya que es el que controla cuanta perturbación se va añadir a cada paso t, el termino $q(x_t|x_{t-1})$ es la condición que describe el estado de la imagen con más ruido a partir de la imagen con menos ruido. Este paso se repite muchas veces, típicamente del orden de mil veces, con el fin de generar una cadena de imágenes cada vez más ruidosas. Se puede mostrar que, cuando el paso crece, la distribución de x_t tiende a parecerse a una **normal estándar**. Este proceso es importante ya que el modelo aprenderá a revertir la imagen con ruido paso a paso, es decir dado un x_t va a predecir como sería x_{t-1} que lo generó desde un inicio, lo cual es importante para generar nuevas imágenes desde ruido puro. No daremos más detalles, pero mencionaremos simplemente que, afortunadamente, en lugar de aplicar este proceso paso por paso, existe una manera simple de generar un dato x_t a partir del dato limpio, x_0, lo que acelera el proceso.

Denoising process

El proceso inverso al que hemos descrito (llamado *Denoising process*) es el que sigue el modelo para poder generar una imagen clara a partir de una imagen con ruido puro y lo va eliminando progresivamente. En otras palabras, lo que trata hacer es predecir $\boldsymbol{x_0}$ (un dato limpio) a partir de $\boldsymbol{x_t}$ mediante varios pasos de tiempo. Se predice una cantidad de ruido $\boldsymbol{\epsilon}$ en $\boldsymbol{x_t}$ en cada paso, de manera condicionada a la imagen de entrada, y así poder obtener $\boldsymbol{x_0}$ de manera gradual.

Este mecanismo que hemos descrito es un **proceso probabilístico** (en este sentido, es un modelo **generativo**, capaz de generar diferentes imágenes plausibles a partir de una realización de ruido). El proceso de *Denoising* se hace a través de una **red neuronal profunda** que predice el ruido agregado, durante el entrenamiento. El modelo minimiza la siguiente función de perdida:

$$\mathcal{L}_{\text{simple}} = E_{\mathbf{x}_0,t,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|^2\right]$$

donde $\boldsymbol{\epsilon_\theta}$ es el ruido que se intenta predecir. Esta red neuronal toma de entrada: la síntesis ruidosa $\boldsymbol{x_t}$para un paso *t,* el paso *t* mismo, y la imagen de la multitud. En salida, se genera una imagen de ruido, de misma dimensión. La arquitectura más usada para este tipo de tareas (imagen de entrada, imagen de salida) es conocida como U-net, la cual en la mayoría de los casos es utilizada en visión por computadora. Notamos que también existen métodos con otras arquitecturas. La **U-net es una red neuronal convolucional** que reduce progresivamente el tamaño de la imagen y extrae las características importantes. Sin embargo, a medida que se baja en la red, se capturan patrones cada vez más abstractos, haciendo que se pierda resolución espacial. En una segunda parte, se procede a pasos en el otro sentido, para producir imágenes de resolución mayor. La siguiente figura (Figura 2) muestra un diagrama del funcionamiento del proceso de difusión en el caso del modelo CrowDiff.

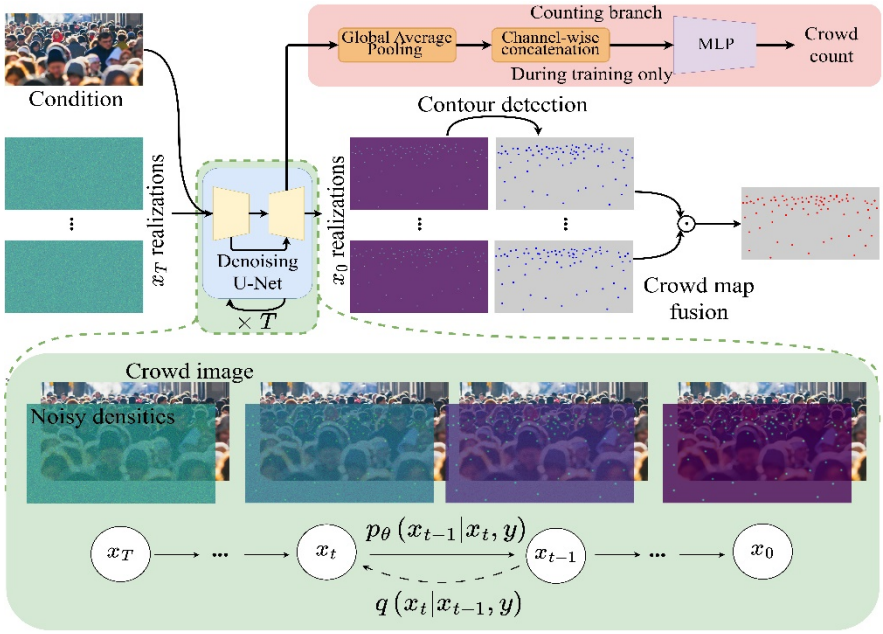


Figura 2. Diagrama explicativo del proceso de difusión en el caso de la red CrowdDiff (Johns Hopkins University, Baltimore, USA).

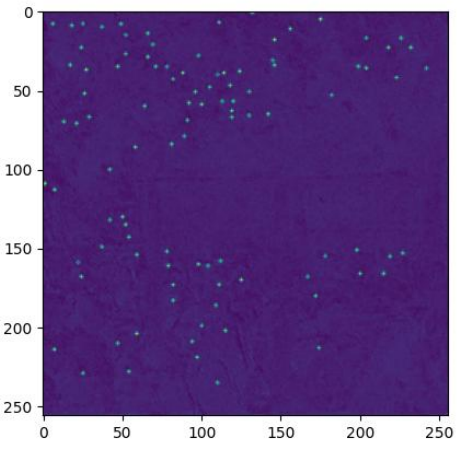
En la figura se puede apreciar como es el "*forward process*" y posteriormente el proceso de "*denoising*" de la imagen cuando entra a la red U-Net, mostrando los pasos intermedios antes de crear los mapas de densidad. En estos pasos intermedios se puede observar el conteo de la multitud a través de puntos de color, que posteriormente van armando el mapa de densidad.

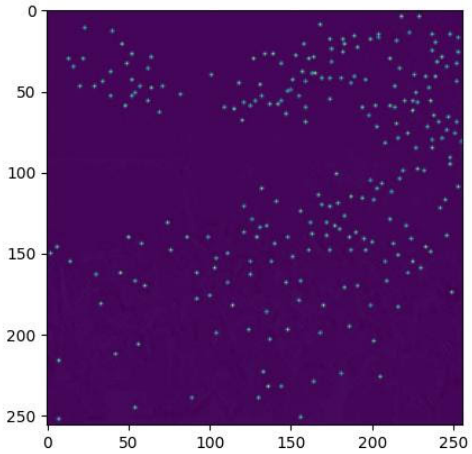
Metodología:

Para desarrollar el modelo, se usó esencialmente Python y la biblioteca Pytorch para entrenar el modelo y realizar los experimentos de test. Cabe notar que existen datasets específicos para este tipo de algoritmos de análisis de multitudes. Para entrenar este modelo, se utilizó el dataset Shangaitech A y Shangaitech B. El primero es un dataset con más carga de población en sus imágenes, mientras que en el segundo es un poco más ligera. Cada uno de estos datasets viene con un "*ground-truth*", el cual es un archivo de formato. mat que sirve para poder entrenar el modelo y poder estimar métricas de error, ya que estos archivos contienen el conteo y la ubicación real de la población presente en la imagen. Las siguientes figuras muestran un ejemplo del conteo de población.

Figura 3. Ejemplo de conteo de multitudes.

La figura 3 muestra cómo trabaja el algoritmo, donde la imagen original del dataset es la primera a la izquierda y las otras tres muestran los procesos de ruido anteriormente explicados en la sección anterior. Estas son imágenes intermedias donde cada punto representa donde el algoritmo cree que hay una persona. Para generar estas imágenes y poder observar los puntos, es necesaria la creación de una función removedora de fondo que trabaje junto con el algoritmo y posteriormente se puede crear otra función adicional para poder realizar el conteo de los puntos. Una vez que se tienen estas imágenes intermedias, es posible generar un mapa de densidad asociado. Las imágenes de la Figura 4 muestran ejemplos de mapas de densidad.





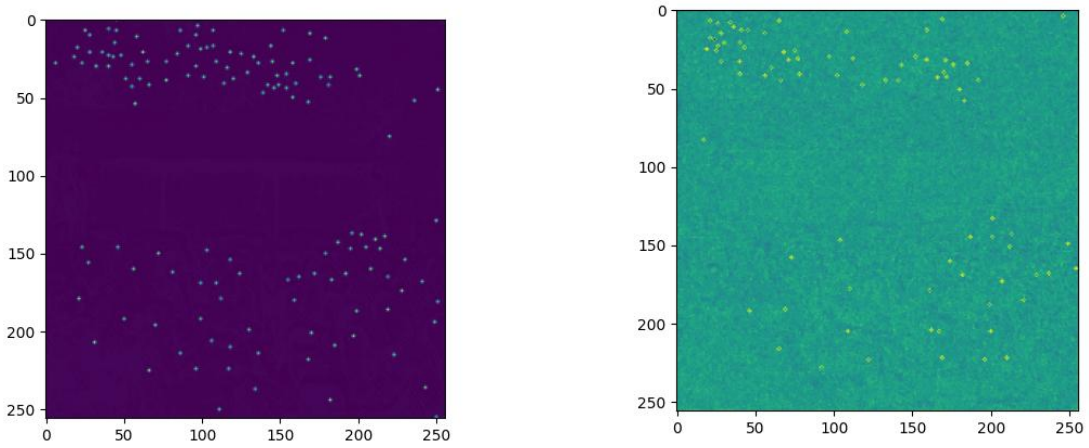


Figura 4. Ejemplos de mapas de densidad.

Sin embargo, también pueden existir algunos problemas de detección por parte del algoritmo y por la naturaleza de las imágenes con las que se entrenó o se está entrenando. Las siguientes figuras muestran algunos ejemplos erróneos en la detección. Posteriormente, se van a explicar algunas sugerencias sobre como poder minimizar este tipo de errores.

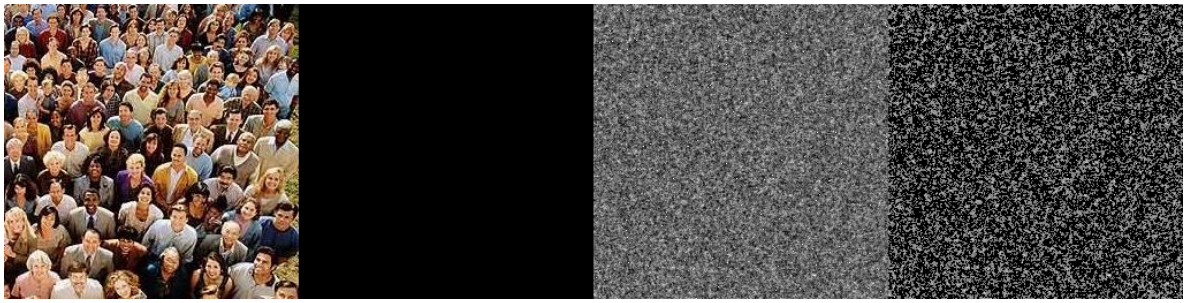


Figura 5. Proceso de "denoising" erroneo.

En la figura 5 se puede observar cómo el proceso de "denoising" no se llevó a cabo correctamente por el algoritmo, generando una imagen con puro ruido. Este tipo de errores son comunes a la hora de implementar el modelo y generalmente tienen que ver con un entrenamiento incompleto o con no tener un modelo con los suficientes pasos para poder realizar el conteo de la multitud.

Resultados:

Como ya se había mencionado anteriormente, una parte de las aplicaciones de este tipo de modelos es para poder diseñar estrategias de seguridad que tengan que ver con la seguridad en eventos con grandes congregaciones de personas. Para ello se analizaron imágenes de una cámara de vigilancia observando el festival de Love Parade en 2010, en Duisburgo, en la cual ocurrió una estampida. Para ello se creó un dataset propio con imágenes extraídas del video y se adaptó el modelo de difusión para poder generar un mapa de densidad. La figura 6 muestra ejemplo de las imágenes utilizadas. Para todos los resultados, se utilizó un modelo de 250,000 pasos de entrenamiento con un mae de 3.15 y un mse de 0.000248. El hardware utilizado se compuso de una tarjeta de video Nvidia Quadro RTX A6000 y se usaron 48Gb de memoria RAM.

Figura 6. Festival de Love Parade.

Sin embargo, debido al tamaño de la imagen, cada una de las imágenes se tuvo que dividir en sub-imágenes (crops) para que tuvieran el tamaño adecuado que el modelo pide (256x256). La figura 7 muestra los resultados obtenidos.

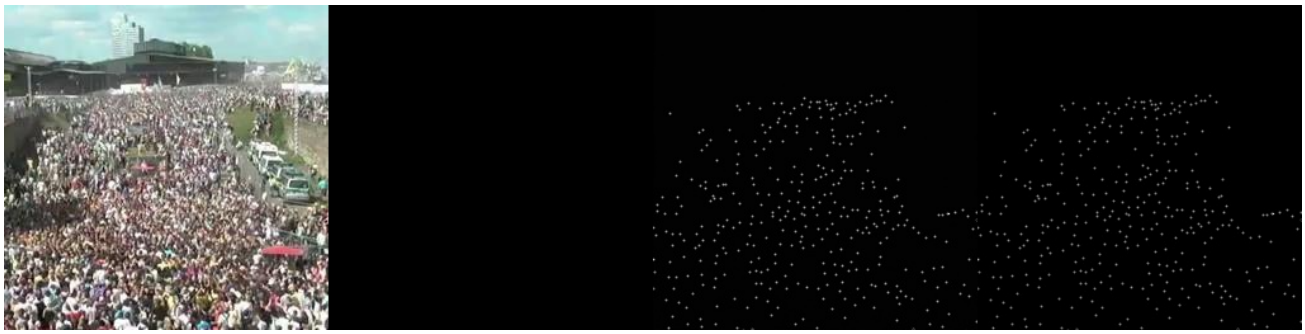


Figura 7. Procesamiento de los mapas de densidad de Love Prade.

De esta manera, se consiguieron los siguientes mapas de densidad. Se muestran en la figura 8

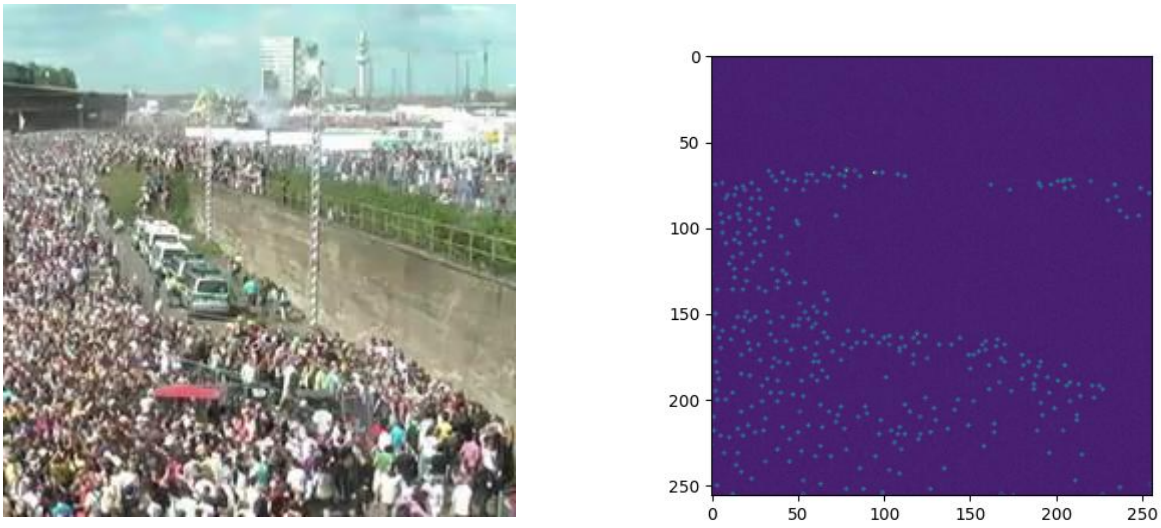


Figura 8 Ejemplo de mapa de densidad de Love Prade.

La figura 9 muestra la imagen reconstruida con los crops utilizados y la figura 10 muestra los mapas de densidad de cada uno.

Figura 9. Crops obtenidos del festival de Love Prade.

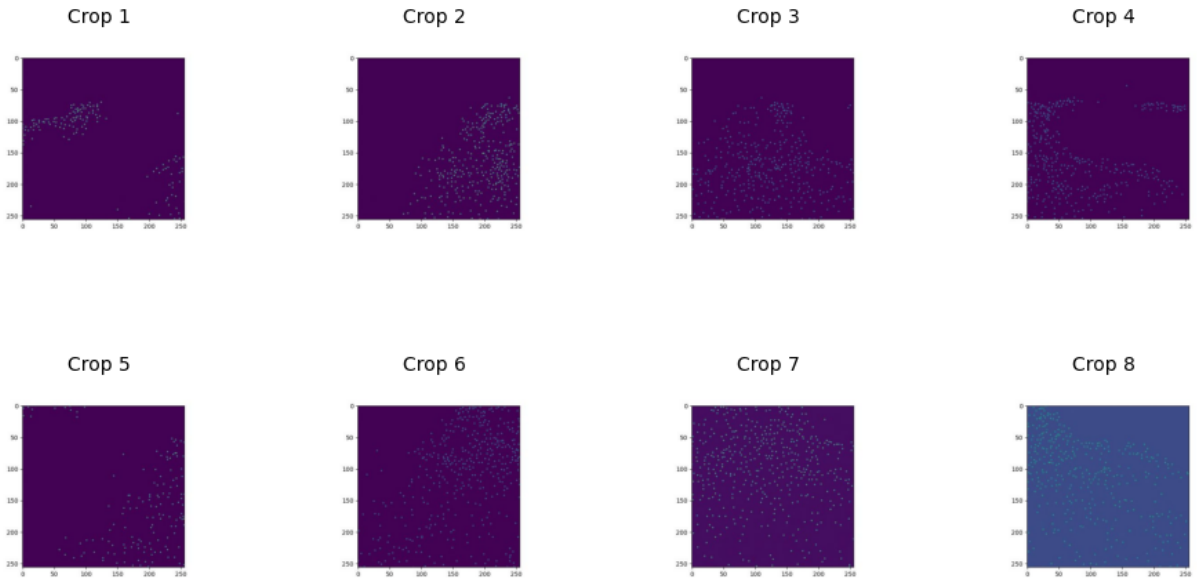


Figura 10. Mapas de densidad de cada crop.

Ahora, un problema común al estar usando imágenes de cámaras de seguridad es la perspectiva que se tiene de la multitud a la que se está observando: Recordamos que la meta que tenemos es producir unos mapas de densidad "métricos" en donde podamos razonar en términos de personas por metro cuadrado. Con los resultados presentados hasta ahora, lo que tenemos son **datos en espacio imagen**, en pixeles, y potencialmente considerablemente deformados con respecto a los datos métricos, por la perspectiva. Para solucionar esto y obtener una estimación real de la densidad de población que había en ese momento en metros cuadrados, se optó por **estimar una homografía** sobre las imágenes usadas con el fin de poder cambiar a una perspectiva más adecuada (de tipo vista de pájaro). De esta manera se puede generar un mapa de densidad más fiel a la geometría de la escena, logrando una estimación general de la cantidad de población cuando la escena observada es plana.

Una **homografía**, es una transformación que asigna los puntos correspondientes de una imagen al punto correspondiente de otra, que existe siempre y cuando la escena observada es plana, lo que es el caso en nuestra situación, ya que la multitud está presente en una gran esplanada. La homografía es una **matriz 3x3** homogénea representada de la siguiente forma:

$$\mathbf{H} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix}$$

Por lo tanto, cualquier par ordenado se produce mediante el siguiente sistema. Es importante notar que para hacer la homografía se seleccionan 4 puntos de la imagen, formando un cuadrilátero y posteriormente a cada uno de esos puntos se les aplica la homografía.

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

Para este caso, se usó la librería OpenCV para la calcular la homografía. La figura 11 muestra un ejemplo de su implementación.

Figura 11. Puntos de entrada.

Una vez seleccionados los puntos específicos, se obtuvo la siguiente salida correspondiente a la homografía de la imagen.

Figura 12. Homografía generada con los puntos de salida.

Ya con la homografía, se hace el mismo procedimiento que se hizo con la imagen original y se obtienen los respectivos mapas de densidad de cada *crop*. La Figura 13 muestra los *crops* que se obtuvieron de la imagen transformada por la homografía. La figura 14 muestra los mapas de densidad obtenidos de cada crop.

Figura 13. Crop de la homografía.

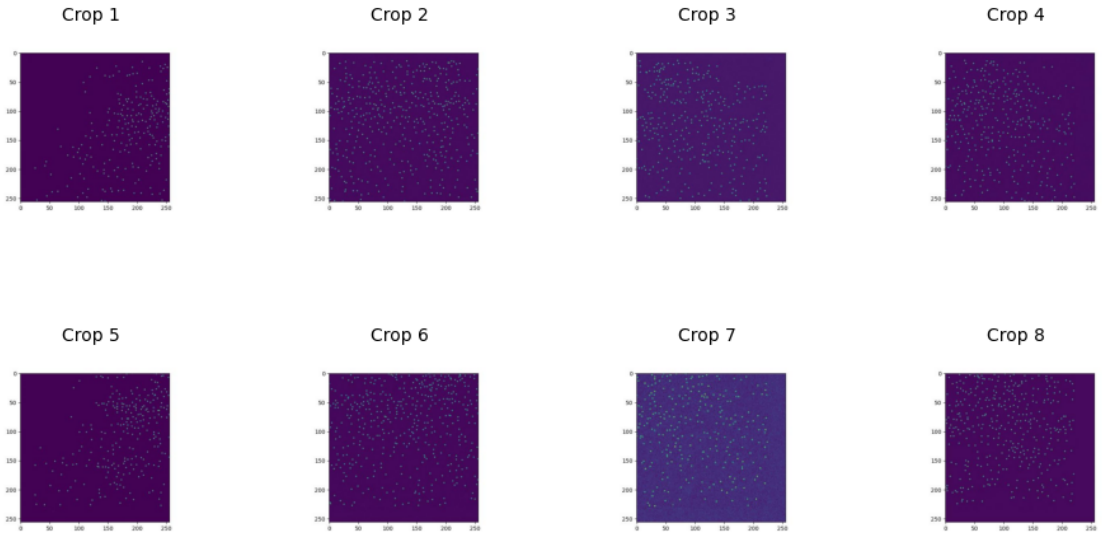


Figura 14. Mapas de densidad de los crops de la homografía.

Mediante el uso de los crops, es más fácil poder generar un mapa de densidad global de la población de la imagen si es que cuanta con medidas más grandes de lo que pide el modelo. Esto también facilita las detecciones en el modelo y hace que los mapas sean más acordes a la cantidad de población real de la imagen.

Conclusiones y trabajo a futuro:

Los modelos de difusión son una herramienta muy poderosa para poder realizar el análisis de multitudes. En este trabajo, hemos mostrado cómo podemos combinar un modelo de estimación de densidad basado en difusión con herramientas geométricas para producir un mapa de densidad. Si se conocen las medidas del lugar en donde se requiere hacer el análisis, hemos visto que es posible estimar una homografía entre la imagen y la escena y expresar la densidad de población del lugar en metros cuadrados. Esto permitiría manejar estos datos de una forma más simple y cuantificable para poder generar estrategias de seguridad o diseño en entornos que requieran una carga de población considerable. Cabe aclarar que la cantidad de recursos computacionales para poder entrenar y testear este tipo de modelos es alta, debido a la cantidad de pasos que se requieren para realizar el entrenamiento. Por lo tanto, mientras mayor sea el tiempo de entrenamiento, aumenta la probabilidad de tener mapas más acordes a las imágenes de entrada.

Como trabajo a futuro, queremos utilizar los datos obtenidos en la representación de mapa de densidad descrita en este documento para entrenar modelos predictivos, que podrían llegar a proveer estimados de la evolución futura del estado de una multitud, y, por ejemplo, nos permitiría desarrollar sistemas de alertas tempranas frente a posibles riesgos de estampidas.

Bibliografía/Referencias

Ranasinghe, Y., Gopalakrishnan Nair, N., Chaminda Bandara, W. G., & Patel, V. M. (2024). CrowdDIfF : Multi-hypothesis crowd density estimation using diffusion models [Articulo]. *Johns Hopkins University, Baltimore, USA*. Recuperado 20 de junio de 2025, de https://openaccess.thecvf.com/content/CVPR2024/papers/Ranasinghe_CrowdDiff_Multi-hypothesis_Crowd_Density_Estimation_using_Diffusion_Models_CVPR_2024_paper.pdf

OpenCV: OpenCV modules. (s. f.). https://docs.opencv.org/4.x/index.html

BBC News. (2020, 4 mayo). *Love Parade disaster: German court ends trial over 2010 stampede deaths*. https://www.bbc.com/news/world-europe-52527515

Dylran. (s. f.). *GitHub - dylran/crowddiff*. GitHub. https://github.com/dylran/crowddiff?tab=readme-ov-file

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, *33*, 6840-6851.

Hartley, R., & Zisserman, A. (2003). *Multiple View Geometry in Computer Vision* (2.ª ed.). Cambridge University Pres.