Universidad de Lima

Facultad de Ingeniería

Carrera de Ingeniería de Sistemas

EVALUACIÓN DEL IMPACTO DE LA SEGMENTACIÓN DE ALETA CAUDAL SOBRE LA FOTO-IDENTIFICACIÓN DE BALLENAS JOROBADAS

Tesis para optar el Título Profesional de Ingeniero de Sistemas

Andrea Jackeline Castro Cabanillas

Código 20160315

Asesor

Oscar Efrain Ramos Ponce

Lima – Perú

Marzo de 2024

Evaluación del impacto de la segmentación de la aleta caudal sobre la foto-identificación de ballenas jorobadas

Castro Cabanillas, Andrea Jackeline
20160315@aloe.ulima.edu.pe
Universidad de Lima

Resumen: La foto-identificación consiste en el análisis de fotografías con el propósito de identificar individuos cetáceos en base a características únicas que cada espécimen de una misma especie exhibe. El uso de esta herramienta permite realizar estudios acerca del tamaño de su población y rutas migratorias mediante la comparación de catálogos. Sin embargo, la cantidad de imágenes que componen estos catálogos es grande, por lo que la ejecución manual de la foto-identificación demanda un tiempo considerable. Por otro lado, muchos de los métodos propuestos para la automatización de esta tarea coinciden en proponer una fase de segmentación para así asegurar que el algoritmo de identificación tome en cuenta únicamente las características del cetáceo y no del fondo. Con el objetivo de determinar si la segmentación mejora los resultados de la identificación de ballenas jorobadas, se construye un método compuesto por una fase de segmentación basado en las redes FCN y PSPNet y una fase de identificación basado en Pérdida de tripletes; se puso a prueba este algoritmo de identificación usando las imágenes sin segmentar y segmentadas. Los resultados indican que la segmentación favorece a la identificación de ballenas jorobadas llegando a un mAP@1 de 0.66 y teniendo el 0.8 de probabilidad de encontrar la identidad correcta cuando los 10 individuos más parecidos son retornados. El método es capaz de clasificar enfocándose en las características del individuo, favoreciendo la identificación especialmente cuando la aleta ocupa una pequeña porción de la imagen original.

Palabras Clave: Inteligencia artificial, Segmentación de imágenes, Visión computacional, Procesamiento de imágenes, Aprendizaje de métricas de distancia, Cetología, Foto-identificación

Abstract: Photo-identification consists of the analysis of photographs with the purpose of identifying individual cetaceans based on unique characteristics that each specimen of the same species exhibits. The use of this tool allows studies of population size and migratory routes by comparing catalogs. However, the number of images that make up these catalogs is large, so the manual execution of photo-identification requires considerable time. Many of the methods proposed for the automation of this task coincide in proposing a segmentation phase to ensure that the identification algorithm considers only the characteristics of the cetacean and not the background. To determine whether segmentation improves humpback whale's identification results, we constructed a method composed of a segmentation phase based on FCN and PSPNet networks and an identification phase based on Triplet Loss; we tested this identification algorithm using unsegmented and segmented images. The results indicate that segmentation favors the identification of humpback whales reaching a mAP@1 of 0.66 and having 0.8 probability of finding the correct identity when the 10 most similar individuals are returned. The method is available to classify by focusing on the characteristics of the individual, favoring identification especially when the whale's fluke occupies a small portion of the original image.

Keywords: Artificial Intelligence, Image Segmentation, Computer Vision, Image Processing, Distance Metric Learning, Cetology, Photo-identification

1. INTRODUCCIÓN

La foto-identificación es el análisis de fotografías con el propósito de reconocer individuos o especímenes de cetáceos. Este análisis se enfoca en la comparación de las características únicas que cada individuo exhibe en distintas partes de su cuerpo; dependiendo de la especie, se podrían tomar en cuenta la pigmentación, las cicatrices y muescas de sus aletas dorsales o caudales, o los distintos patrones de callosidades en sus cabezas. Para los biólogos marinos, la foto-identificación es una herramienta importante, ya que permite realizar estudios sobre la vida silvestre; por ejemplo, se pueden determinar los patrones de migración de una población mediante la comparación de catálogos de fotografías de los individuos vistos en diferentes lugares (Titova et al, 2018); también, constituye un método indispensable para estimar el tamaño de una población teniendo en cuenta el número de individuos avistados en una zona durante distintas temporadas de tiempo (Barlow et al., 2010; Félix et al., 2011); finalmente, mediante la foto-identificación se puede llevar un registro de la estructura del grupo poblacional, así como la fidelidad al punto geográfico (Ballance, 2018).

Las ballenas jorobadas son conocidas por realizar la migración más extensa con distancias de entre 6 y 8 mil kilómetros cada año (Ruiz, 2016) y por deleitar con sus vistosas acrobacias en el aire durante su paso. Estas acrobacias aéreas las convierten en la principal atracción durante la actividad turística del avistamiento de ballenas, que se realiza en múltiples países del mundo como Canadá, Estados Unidos, Sudáfrica, Nueva Zelanda, Hawái y Perú, contribuyendo a su economía. En el Perú, 45 mil turistas participaron en la actividad de avistamiento de ballenas

jorobadas al cierre del 2019 (Barrientos, 2019); además, para el año 2023 se espera que esta actividad turística genere impacto económico de 18 millones de soles en el norte del Perú (Gestión, 2023). Gracias a los estudios basados en la foto-identificación, se puede conocer más sobre un grupo de ballenas jorobadas llamado Stock G que llega al Perú; estas investigaciones han determinado que la población llega durante el invierno al norte del Perú, para reproducirse y dar a luz a sus crías, y parte desde la Antártida donde se abastece de alimento durante el verano (Ruiz, 2016); además, se ha estimado que el Stock G está compuesto por 6504 individuos (Félix et al, 2011); pero, un estudio reciente realizado en 2017 estima una tasa de crecimiento poblacional del 2,31% anual, una cifra preocupante debido a que en 2011 se indicó una tasa de crecimiento anual del 6,3% (Monnahan, 2019).

Tradicionalmente, la foto-identificación se realiza de forma manual; sin embargo, el proceso manual exige un esfuerzo considerable ya que implica la revisión de extensos catálogos de imágenes que pueden llegar a ser miles de fotografías (Maglietta et al., 2018; Weideman et al., 2017). Además, Weideman et al. (2017) afirma que los cambios de pose del animal, la variante del punto de vista y las partes a veces ocultas por el mar dificultan esta tarea. Por estas razones, los estudios basados en la foto-identificación no suelen extenderse a una población mayor, o a periodos de tiempo más largos (Bouma et al., 2018).

En el estado del arte se encuentran métodos que pretenden automatizar la foto-identificación de diferentes especies de cetáceos, como el delfín de Risso, el delfín común, la ballena franca y la ballena jorobada, utilizando técnicas de extracción de características, aprendizaje profundo tradicional y aprendizaje métrico a distancia. Los autores informan de dificultades como un pequeño número de imágenes por individuo, diferencias muy sutiles entre individuos y fondos muy similares, compuestos por mar y cielo, que pueden llevar al algoritmo a una identificación errónea (Hsu et al., 2018; Pollicelli et al., 2020). Por este motivo, algunos trabajos añaden una fase de segmentación previa a la fotoidentificación para asegurar que el algoritmo seleccionado se basa únicamente en las características del individuo a la hora de emparejarlo y no en su entorno (Gilman et al., 2016; Hsu et al., 2018; Maglietta et al., 2018; Reno et al., 2018; Weideman et al., 2017). Sin embargo, no se han encontrado investigaciones que prueben si la segmentación previa devuelve mejores resultados de identificación en ballenas jorobadas.

En la presente investigación, se pretende probar si la segmentación mejora el rendimiento del algoritmo de foto-identificación en la especie de ballena jorobada. Con este objetivo, se implementará un algoritmo de segmentación híbrido compuesto por las técnicas que obtuvieron los mejores resultados en Castro & Ayma (2021): FCN y PSPNet. Además, se construirá el algoritmo de identificación basado en el aprendizaje de métricas de distancias y se probará sobre la base de datos Humpback Whale Identification Challenge (Kaggle, 2018), tanto segmentada como sin segmentar, para comprobar la hipótesis.

El resto de este trabajo se organiza de la siguiente manera, la sección 2 presenta el estado del arte asociado a las propuestas de algoritmos de foto-identificación en distintas especies de cetáceos, la sección 3 describe las técnicas utilizadas en nuestra implementación; la sección 4 presenta la metodología propuesta, así como los detalles relativos a la implementación; finalmente, La sección 5 explica los resultados obtenidos y la sección 6 muestra las conclusiones.

2. ESTADO DEL ARTE

La foto-identificación es una técnica frecuentemente utilizada por los biólogos marinos para estimar el tamaño de las poblaciones de ballenas jorobadas, determinar sus patrones migratorios y monitorear el estado de grupos de interés en particular. A pesar de la importancia de estos estudios para la conservación de la especie, existen pocos trabajos orientados a la automatización de esta tarea sobre ballenas jorobadas; por esta razón, la revisión también incluirá trabajos relacionados con la foto-identificación en otras especies de cetáceos.

La revisión comenzará describiendo aquellos trabajos que se enfocan en el desarrollo de métodos de identificación de ballenas y cetáceos; posteriormente, se describirán investigaciones que incluyen al aislamiento del área de interés como una etapa previa al reconocimiento de ballenas y cetáceos.

A. Reconocimiento de ballenas y cetáceos

La foto-identificación de cetáceos ha sido abordada bajo diversas perspectivas, entre ellas el procesamiento de imágenes y el aprendizaje profundo. En el primer caso, la foto-identificación se consigue al combinar técnicas tradicionales de extracción y correspondencia de características para identificar individuos cetáceos. Por ejemplo, Joly et al. (2016) propusieron emplear el algoritmo de RANSAC para identificar ballenas jorobadas a través del descubrimiento de biomarcadores comunes en pares de imágenes de aletas caudales de las ballenas. Los biomarcadores son codificados en vectores de características producto de la aplicación del algoritmo de procesamiento de imágenes SIFT. Esta propuesta alcanza una precisión de reconocimiento de alrededor 0.49. De acuerdo a los autores, este resultado podría mejorar si el algoritmo SIFT se ejecutase sobre la región de la imagen correspondiente a la aleta caudal que contiene los biomarcadores de la ballena jorobada.

A diferencia de las técnicas de procesamiento de imágenes, las técnicas de aprendizaje profundo tradicionales se centran en aprender conjuntamente a extraer las características más distintivas de los individuos cetáceos y a determinar la correspondencia subsecuente. Por un lado, Gómez Blas (2020) propusieron el despliegue de una red neuronal convolucional (CNN) para identificar individuos de ballena jorobada, que se compone de dos capas convolucionales y una red neuronal multicapa con 4251 neuronas de salida equivalentes al número individuos a identificar. Este método solamente alcanza una precisión del 0.44, siendo el principal inconveniente el desbalance en la cantidad de imágenes por individuo en este tipo de cetáceos. Para subsanar este inconveniente, los autores proponen como trabajo futuro utilizar aprendizaje de métricas de distancia. Este enfoque de aprendizaje profundo busca predecir la similaridad o no similaridad entre dos imágenes, otorgándole la identidad del más parecido. De esta manera, la red neuronal convolucional (CNN) obtiene más ejemplos durante el entrenamiento al multiplicarse las combinaciones posibles de imágenes similares y no similares de las cuales puede aprender.

Una red neuronal convolucional (CNN) entrenada con el enfoque aprendizaje de métricas de distancia, otorga una representación robusta a cada individuo; es decir, un vector numérico que al ser comparado con los demás vectores pueda otorgarse la identidad del más parecido. Por ejemplo, Schneider et al. (2020) compararon el desempeño en la foto-identificación de ballenas jorobadas a partir de arquitecturas de redes convolucionales conocidas, entre ellas la AlexNet, VGG-19, DenseNet201, MobileNetV2 e InceptionV3, pero utilizando el enfoque de aprendizaje de métricas de distancia con Redes Siamesas y Pérdida de tripletes. Los resultados de este estudio señalaron que el mejor desempeño fue alcanzado por la arquitectura InceptionV3 entrenada con pérdida de tripletes obteniendo un 0.746 de mAP.

Estos métodos propuestos se concentran en proponer métodos automatizados de identificación de ballenas jorobadas; sin embargo, no incluyen alguna etapa de aislamiento de la aleta del cetáceo previa para poder asegurarse de que la foto-identificación se realiza en base a las características del individuo y no de su entorno.

B. Aislamiento de aletas para el reconocimiento de ballenas y cetáceos

En vista de que la información proveniente del entorno en las imágenes de cetáceos podría estar incluyéndose de manera errónea en la foto-identificación, existen trabajos de investigación que incluyen una etapa de aislamiento previa a la identificación, el cual se puede observar de distintas maneras: mediante la detección del contorno de la aleta, el recorte del área de interés dentro de la imagen o la segmentación de la región de la aleta del cetáceo. Estas diversas maneras de aislamiento cuentan con el objetivo de lograr una identificación en base a las características del individuo mas no de su entorno y, al igual que en la sección anterior, tanto el aislamiento como identificación puede ser abordada desde el área de estudio de procesamiento de imágenes, aprendizaje automático o aprendizaje profundo.

El aislamiento de las regiones de interés para promover un mejor reconocimiento de cetáceos tuvo inicio con el trabajo propuesto por Ramos-Arredondo et al. (2014), quienes emplearon el algoritmo *Active Contours* para detectar el contorno de la aleta perteneciente a las ballenas azules, posteriormente, utiliza la técnica de Momentos de Hu para características invariables en escala, traslación y rotación que identifiquen a cada individuo; por último, utiliza el algoritmo K vecinos más cercanos (K-NN) para lograr la clasificación entre estos tres tipos de aletas, alcanzando un 0.45 de precisión. De forma similar, Gilman et al. (2016) emplearon el algoritmo el algoritmo *Iterative Closest Point* para detectar el contorno de las aletas de delfines comunes; posteriormente, extrae indicadores estadísticos, como desviación estándar y la media del valor de sus píxeles, de subdivisiones de la aleta; finalmente, utiliza un análisis de la discriminante lineal (LDA), un clasificador probabilístico, para realizar la identificación y alcanzan un 0.71 de precisión.

En vista a los resultados prometedores obtenidos por Gilman et al. (2016) en el reconocimiento de delfines en base a los contornos de sus aletas, Weideman et al. (2017) y Thompson et al. (2019) proponen emplear algoritmos de aprendizaje profundo para complementar la detección de contornos con capacidad representativa de las CNN. Bajo esta premisa, Weideman et al. (2017) utiliza una Red Neuronal Completamente Convolucional (FCN) para aislar el contorno de la aleta de delfines cuello de botella y ballenas jorobadas, luego utiliza un algoritmo de deformación dinámica en el tiempo para alinear una representación del contorno no conocida sobre los demás ya conocidos, el individuo más similar será aquel que cuente con un menor costo de alineamiento; es decir, una menor suma de errores. Este método alcanza una precisión de 0.74 en delfines y 0.86 en ballenas. Mientras tanto, Thompson et al. (2019) construye dos redes convolucionales, ambas basándose en la arquitectura ResNet. Con la primera busca identificar el borde de la aleta dorsal de delfines cuello de botella en sus respectivas fotografías, posteriormente, utiliza únicamente el contorno de la aleta para entrenar una segunda arquitectura y así pueda identificar a cada individuo. Esta propuesta logra un 0.97 de precisión al retornar los 50 individuos más parecidos. Hacer uso del contorno de la aleta del individuo para la identificación presenta buenos resultados en los trabajos mencionados; sin embargo, podría existir un

desaprovechamiento en la información de los patrones de pigmentación que guardan cada cetáceo dentro de su aleta.

Por otro lado, algunos autores proponen recortar la región rectangular de la imagen que enmarca la aleta del cetáceo. Bouma et al. (2018) propuso utilizar la arquitectura GoogleLeNet para detectar y recortar la aleta dorsal de delfines comunes de la imagen. Debido a la variedad de individuos, los autores aplican aumento de datos a partir de cambios en la tonalidad, saturación y rotación de las imágenes originales. Posteriormente, para la identificación utiliza la arquitectura ResNet50 la cual es inicializada con pesos pre-entrenados en sobre el conjunto de datos ImageNet. Esta propuesta alcanza un 0.90 de precisión. Mientras tanto, Bogucki et al. (2018) propuso el uso de tres redes convolucionales que se encargaran de tareas distintas. La primera es entrenada para localizar la cabeza de la ballena franca en la fotografía; posteriormente, entrena otra red para identificar los puntos de inicio y fin de la callosidad, con la finalidad de rotar y recortar las imágenes de manera estandarizada; finalmente, utiliza una última red convolucional para la clasificación. Esta propuesta logra una precisión de 0.87. Utilizar el recorte de la aleta del cetáceo para la identificación resulta conveniente ya que minimiza la posibilidad de perder información útil al redimensionar las imágenes que pasan por el algoritmo de aprendizaje profundo; sin embargo, estas imágenes aún contarán con una porción del entorno.

Además, existen trabajos que optan por la segmentación de la región de la aleta del cetáceo como etapa de aislamiento. Por ejemplo, Reno et al. (2018) y Maglietta et al. (2018) emplearon técnicas tradicionales de procesamiento de imágenes para el aislamiento y reconocimiento de los individuos de delfín común y delfines de Risso, respectivamente. Reno et al. (2018) hace uso del algoritmo *Speeded Up Robust Features* (SURF) sobre la imagen de aleta de delfín segmentada con Otsu para localizar características distintivas en ella. La comparación entre las características de aleta de un delfín candidato con aquellas pertenecientes a otros individuos en un catálogo permitió identificar correctamente al 89% de los individuos. Mientras tanto, Maglietta et al. (2018), utiliza el algoritmo de Otsu para la segmentación automática de la aleta de delfines de Risso; a continuación, utiliza el algoritmo SIFT para extracción de características distintivas de las aletas y, finalmente, emplea el algoritmo de aprendizaje profundo RUSBoost para la clasificación de los individuos, alcanzando un 0.84 de precisión. Por otro lado, Hsu et al. (2018) utilizan un método híbrido basado en mapas se prominencia para segmentar la imagen de la cola de delfines comunes antes de que pase al modelo de clasificación basado en DenseNet121; de esta manera, logra obtener un 0.85 de precisión; adicionalmente, Hsu et al. (2018) realizan un experimento en donde compara el rendimiento del algoritmo de identificación sin su etapa de aislamiento y obtiene un 0.4 de precisión; entonces, concluye que el modelo podría clasificar basándose en la apariencia del mar y no de los delfines. Este último enfoque de aislamiento se diferencia de los anteriores ya que se aprovecha la información tanto del contorno como la pigmentación que tienen las aletas de los cetáceos; además, busca asegurarse de remover la información de su entorno por completo antes de pasar por una etapa de identificación. De esta manera, aprovecha toda la información del individuo dejando de lado la información del entorno.

Adicionalmente, Castro & Ayma (2021) evaluaron el desempeño de distintos algoritmos en la tarea de segmentación de aletas de ballena jorobada. Entre los algoritmos se encontraron dos técnicas de procesamiento de imágenes: Otsu y Chan Vese, y dos algoritmos de visión computacional: las redes FCN y PSPNet. Concluyeron que las redes FCN y PSPNet, con un IoU promedio de 0.9434 y 0.9433, respectivamente, tienen un mejor desempeño que los algoritmos de procesamiento de imágenes; además, analizaron que sus resultados son complementarios.

Finalmente, en base a la revisión del estado del arte, se decidió implementar un algoritmo de segmentación con las redes FCN y PSPNet ya que obtuvieron los mejores resultados en Castro & Ayma (2021) en la tarea de segmentación de aletas de ballena jorobada; por otro lado, para la fase de identificación, se optó por utilizar el enfoque de aprendizaje de métricas de distancia ya que obtuvo buenos resultados en Schneider et al. (2020) en la tarea de identificación de ballenas jorobadas y sobre el conjunto de datos más amplio del estado del arte; además, para poner en práctica este enfoque, se utilizará la red Resnet-50 para la construcción de representaciones ya que tuvo un buen desempeño en Bouma et al. (2018) y por la sencillez de su entrenamiento debido a que está compuesta por una menor cantidad de capas a comparación de otras arquitecturas; y al algoritmo KNN debido a que Schroff et al. (2015), quienes propusieron este enfoque de aprendizaje profundo, recomiendan utilizarlo para la última etapa de clasificación en base a representaciones.

3. ANTECEDENTES

En esta sección se presentan los fundamentos teóricos con relación al algoritmo propuesto: Redes neuronales convolucionales (CNN), segmentación semántica y la teoría detrás del aprendizaje de métricas de distancia.

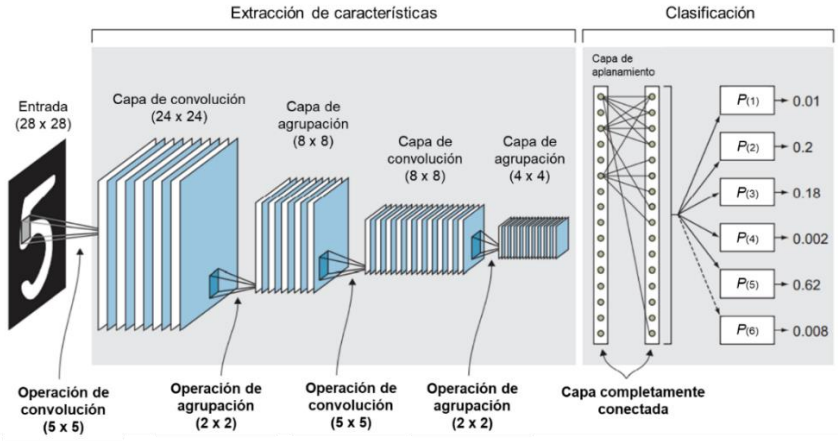
3.1 Redes neuronales convolucionales (CNN)

Las redes neuronales convolucionales (CNN) son un tipo de redes neuronales especializadas en tareas de visión computacional como la clasificación de imágenes, detección de objetos, búsqueda de imágenes, entre otras. Una

arquitectura de CNN toma como entrada una imagen, utiliza múltiples capas de convolución y de agrupación para poder extraer sus características; finalmente, hace uso de capas completamente conectadas para clasificar satisfactoriamente a la imagen. En la Fig. 3.1, se puede observar un ejemplo de arquitectura de CNN; a continuación, se presentan los componentes mencionados:

Figura 3.1

Arquitectura de CNN



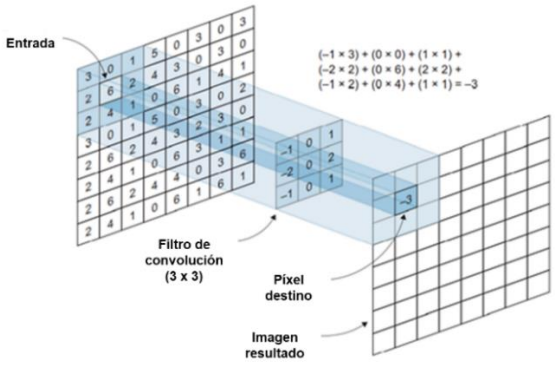
Nota. Adaptado de Elgendy (2020)

3.1.1 Capa de convolución

Las capas de convolución tienen el objetivo de detectar características predominantes en las imágenes. Para ello, recorre todos los píxeles de la imagen realizando una multiplicación de matrices entre esta y un filtro de convolución, el cual actuará como detector de características. Los valores del filtro de convolución son aprendidos durante la etapa de entrenamiento de la CNN. Por ejemplo, en la Fig. 3.2 se puede observar el funcionamiento de un filtro de convolución de tamaño 3 x 3 píxeles, el cual recorre toda la imagen de tamaño 8 x 8 y retorna una matriz de las mismas dimensiones con el resultado de la multiplicación matricial que permitirá obtener las características distintivas.

Figura 3.2

Capa de convolución



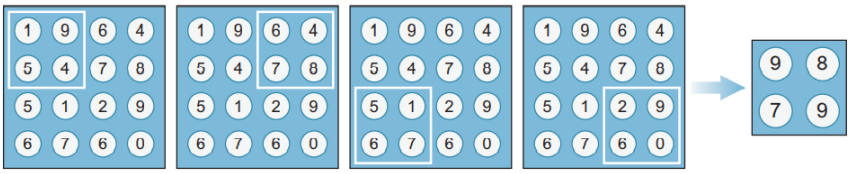
Nota. Adaptado de Elgendy (2020)

3.1.2. Capa de agrupación

Las capas de agrupación tienen como objetivo reducir el número de parámetros en la arquitectura de CNN. Para ello, aplica operaciones estadísticas, de promedio o máximo valor, por todos los píxeles de la imagen. Por ejemplo, en la Fig. 3.3 se puede observar el funcionamiento de un filtro de agrupación de máximo valor de tamaño 2 x 2 píxeles, el cual recorre toda la imagen de tamaño 4 x 4 píxeles y retorna una matriz con una menor dimensión de 2 x 2 píxeles; permitiendo que las siguientes capas contengan menos parámetros al mismo tiempo que manteniendo las características más importantes.

Figura 3.3

Capa de agrupación



Nota. De Elgendy (2020)

3.1.3. Capa completamente conectada

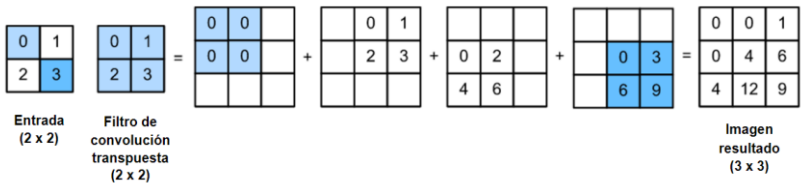
Por último, la capa completamente conectada tiene como objetivo lograr la clasificación final de la imagen. Primero, hace uso de una capa de aplanamiento, la cual se encarga de aplanar la matriz resultado de las capas de convolución y agrupamiento en un vector de características; posteriormente; puede añadir una o más capas conectadas intermedias para mejorar el aprendizaje antes de la clasificación en la capa de salida. Por ejemplo, en la Fig. 3.1, esta capa retorna la probabilidad de la imagen de pertenecer a cada una de las clases involucradas.

3.1.4. Capa de convolución transpuesta

Las capas de convolución transpuesta tienen el objetivo de incrementar la resolución de la matriz de salida. En la operación de convolución traspuesta un filtro de convolución recorre toda la imagen multiplicando los valores entre las matrices; sin embargo, para lograr una salida de mayor resolución los resultados van acumulándose en una matriz de mayor tamaño para sumarse al final. Por ejemplo, en la Fig. 3.4 se muestra la operación de convolución traspuesta de un filtro de tamaño 2 x 2 píxeles sobre una imagen de entrada de 2 x 2 píxeles, el cual resulta en una imagen de 3 x 3 píxeles.

Figura 3.4

Operación de convolución traspuesta



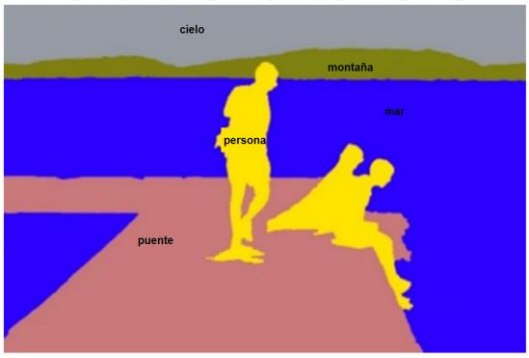
Nota. Basado en Elgendy (2020)

3.2. Segmentación semántica

La segmentación semántica busca predecir la clase de cada píxel en una imagen. Por ejemplo, en la Fig. 3.5, la predicción ayuda a determinar la clase a la cual pertenece cada píxel; es decir, si este pertenece a una persona, a un puente, al cielo, a una montaña o al mar.

Figura 3.5

Ejemplo de segmentación semántica



Nota. De Shanmugamani (2018)

A diferencia de las tareas de clasificación, en las que se devuelve una sola etiqueta, las arquitecturas de redes

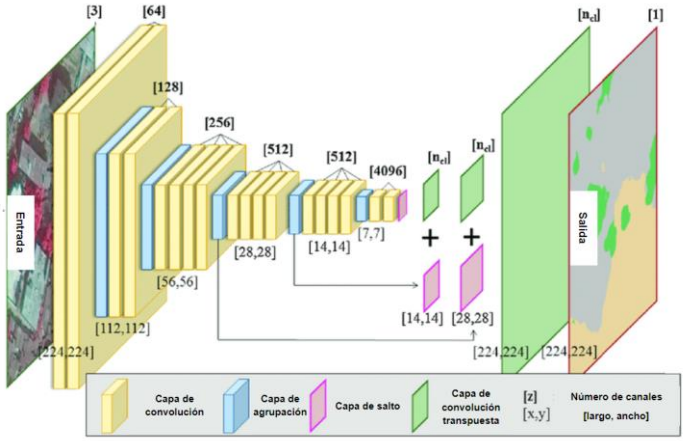
neuronales convolucionales (CNN) diseñadas para la segmentación semántica generan una imagen de salida que mantiene la misma resolución que la entrada, asignando etiquetas de clase para cada píxel. A continuación, se presentan las arquitecturas de Red completamente convolucional (FCN) y Red de análisis de escenas en pirámide (PSPNet) que obtuvieron los mejores resultados en Castro & Ayma (2021), donde fueron evaluadas en la segmentación aletas de ballenas jorobadas, y se utilizaron en la metodología propuesta.

3.2.2. Red completamente convolucional (FCN)

La red completamente convolucional (FCN), propuesta por Long et al. (2015), está especializada en tareas de segmentación de imágenes. Su arquitectura se compone por capas de convolución y de agrupación de promedio; además, se distingue de las demás redes ya que incorpora capas de salto, las cuales fusionan información de capas convolucionales en distinto nivel; es decir, combinan la información semántica de las últimas capas convolucionales con la información espacial que tienen las capas iniciales la cual tiende a perderse; de esta manera, mejora la eficacia de la segmentación; finalmente, se utilizan capas de convolución transpuesta para generar un mapa de calor con la misma resolución de la imagen de entrada y que contiene a las etiquetas de cada píxel. A continuación, en Fig. 3.6, se muestra la arquitectura de la red completamente convolucional (FCN).

Figura 3.6

Arquitectura de red completamente convolucional (FCN)



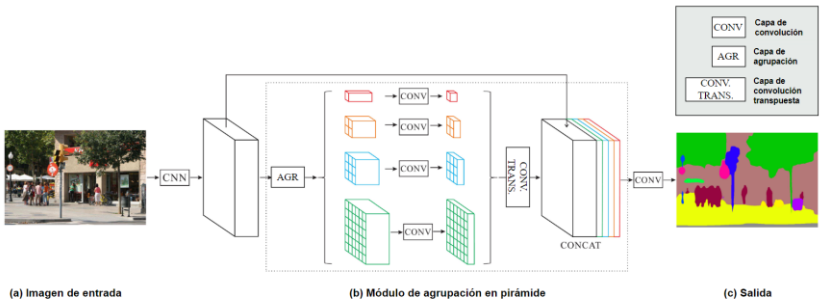
Nota. De Wurm et al. (2019)

3.2.3. Red de análisis de escenas en pirámide (PSPNet)

La red de análisis de escenas en pirámide (PSPNet), propuesta por Chen et al. (2018), también se encuentra especializada en tareas de segmentación semántica. Su arquitectura se diferencia de las demás redes ya que incorpora un módulo de agrupación en pirámide la cual, a través de operaciones de convolución y agrupamiento en distintas escalas, permite evitar la pérdida de información espacial en distintas escalas. Además, utiliza capas de convolución dilatada las cuales, a diferencia de las tradicionales, añaden ceros entre sus valores del filtro con el mismo objetivo de preservar información espacial necesaria para la clasificación a nivel de píxel; finalmente, se utiliza capas de convolución transpuesta para retornar una imagen con las mismas dimensiones que la imagen de entrada. A continuación, en Fig. 3.7, se muestra la arquitectura de la red de análisis de escenas en pirámide (PSPNet).

Figura 3.7

Arquitectura de red de análisis de escenas en pirámide (PSPNet)



Nota. De Chen et al. (2018)

3.3. Aprendizaje profundo de métricas

El aprendizaje profundo de métricas, o aprendizaje de similitud, tiene como objetivo construir una función de similitud que permita proyectar imágenes de entrada (I) en un espacio nuevo (E) de tal forma que las proyecciones de las imágenes originales con características similares se encuentren en ubicaciones próximas en el espacio E; mientras que las menos similares se encuentren alejadas entre sí. De esta manera, la distancia entre dos imágenes permitirá medir la similitud entre ambas (Ge et al., 2018). Este enfoque ha tenido éxito en varias tareas de visión por computador como la verificación de rostros (Parkhi et al., 2015; Schroff et al., 2015), re-identificación de personas (Shi et al., 2016; Ustinova et al., 2016), búsqueda de imágenes (Song et al., 2016; Wohlhart et al., 2015), seguimiento de objetos (Tao et al., 2016), entre otros.

Hoy en día, la aparición de las Redes Neuronales Convolucionales (CNNs) ha permitido que la función de similitud proyecte las imágenes de entrada sobre el espacio E en base a sus características más relevantes y distintivas. De acuerdo con Elgendy (2020), los elementos del aprendizaje profundo de métricas se constituyen por las representaciones, la función de pérdida y la minería de tripletes; además, el entrenamiento de los modelos Red residual número 50 (Resnet-50) y K vecinos más cercanos (KNN) permitieron poner en práctica aprendizaje profundo de métricas en la metodología propuesta. A continuación, se presentan los conceptos mencionados.

3.3.2. Representación

Como se puede observar en la Fig. 3.8, una representación (x) se trata de un vector numérico de *N* dimensiones el cual resulta de la aplicación de la función de similitud sobre una imagen de entrada *I*. Es deseable que la representación preserve las características distintivas de la imagen de entrada ante cambios en el punto de vista, iluminación y pose del objeto. Bajo esta premisa las representaciones de las imágenes más similares se ubican muy próximas en el espacio E. En la actualidad, se puede utilizar cualquier arquitectura de Redes Neuronales Convolucionales para la construcción de las representaciones en conjunto con métricas de distancia ya conocidas, como la distancia Euclidiana o Manhattan (Schroff et al., 2015).

Figura 3.8

Construcción de la representación de una imagen

Nota. De Schroff (2015)

3.3.3. Función de pérdida

Para que una CNNs aprenda a construir representaciones de imágenes se requiere la optimización de sus parámetros de forma que estos minimicen un valor de error entre distancias, el cual se define en la función de pérdida. Actualmente, la función de pérdidas que más destaca en las tareas de identificación es la función de pérdida de tripletes, la cual considera tres elementos en su formulación: ejemplos ancla (x_a), ejemplos negativos (x_n), y ejemplos positivos (x_p). Un ejemplo ancla (x_a) es la representación de una imagen (I_a) que sirve como referencia para comparaciones de similitud subsecuentes; un ejemplo negativo (x_n) es la representación de una imagen (I_n) disimilar y de identidad distinta a la imagen de referencia; finalmente, un ejemplo positivo (x_p) es la representación de la imagen (I_p) que es similar a la imagen de referencia y por tal comparte la misma identidad que I_a. Entonces, la función de pérdida de tripletes,

$$L = \max\left(D(x_a, x_p) - D(x_a, x_n) + \alpha, 0\right) \quad (1)$$

buscará que la distancia Euclidiana $\boldsymbol{D}$ entre $\boldsymbol{x_a}$ y $\boldsymbol{x_p}$ sea menor a la distancia euclidiana entre $\boldsymbol{x_a}$ y $\boldsymbol{x_n}$; además el valor de margen $\boldsymbol{\alpha}$ evitará que las representaciones estén conformadas por ceros (Schroff et al., 2015).

3.3.4. Minería de tripletes

El uso de CNNs que empleen la función de pérdida de tripletes con el objetivo de producir representaciones de las imágenes que reciba, implica un entrenamiento mediante la recepción de tripletes de imágenes, que son conjuntos de imágenes conformados por un ejemplo ancla, un ejemplo positivo y un ejemplo negativo. Si bien, la selección de tripletes podría ser aleatoria; lo recomendable es seguir una estrategia de minería de tripletes para conseguir mejores resultados (Schroff et al., 2015).

Existen tres tipos de tripletes dependiendo de la posición del ejemplo negativo en el espacio euclidiano:

- **Tripletes simples:** Son aquellos que tienen un valor de pérdida de 0, cumpliendo con la siguiente condición:

$$D(x_a, x_p) + \alpha < D(x_a, x_n) \quad (2)$$

donde, la distancia euclidiana D es menor entre la representación del ejemplo ancla x_a y del ejemplo positivo x_p y mayor entre la representación del ejemplo ancla x_a y del ejemplo negativo x_n.

- **Tripletes complejos:** Son aquellos que cumplen con la siguiente condición:

$$D(x_a, x_n) < D(x_a, x_p) \quad (3)$$

es decir, en estos casos la representación del ejemplo negativo x_n se encuentra más cerca de la representación del ejemplo ancla x_a que del ejemplo positivo x_p.

- **Tripletes de mediana complejidad:** Son aquellos que cumplen con la siguiente condición:

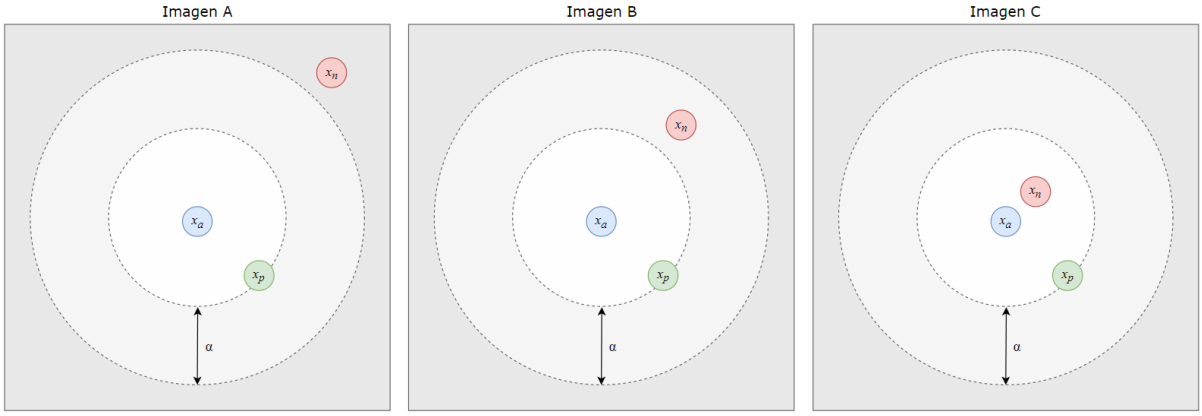
$$D(x_a, x_p) < D(x_a, x_n) < D(x_a, x_p) + \alpha \quad (4)$$

donde, la distancia euclidiana D sí cumple con ser menor entre la representación del ejemplo ancla x_a y del ejemplo positivo x_p, y mayor entre la representación del ejemplo ancla x_a y del ejemplo negativo x_n; sin embargo, la diferencia entre distancias aun no sobrepasa el valor de margen α.

En la Fig. 3.9, se puede observar ejemplos de los tipos de tripletes: en la Imagen A se muestra un triplete simple; en la Imagen B, un triplete medianamente complejo; y en la Imagen C, un triplete complejo.

Figura 3.9

Tipos de tripletes



Nota. De Schroff (2015)

En los casos en los que se emplee CNNs con funciones de pérdidas basadas en tripletes para tareas de identificación, Schroff et al. (2015) recomienda realizar un entrenamiento de las CNNs seleccionando tripletes de mediana complejidad o complejos por cada lote de imágenes; es decir, con aquellos que aún tienen un valor de pérdida mayor a 0; de esta manera, solo estos influyen en la actualización de parámetros en la red neuronal y se lograrían mejores resultados.

3.3.5. Red residual número 50 (Resnet-50)

La red residual número 50 (Resnet-50) es una arquitectura de red neuronal convolucional propuesta por He et al. (2015), esta red se encuentra compuesta por 50 capas: 48 capas de convolución y 2 capas de agrupación; además, incorpora conexiones residuales las cuales buscan preservar la información de las capas iniciales para lograr mejores predicciones. En la metodología propuesta, se entrenó esta arquitectura bajo el paradigma de aprendizaje profundo de métricas para poder obtener las representaciones de las imágenes.

3.3.6. K vecinos más cercanos (KNN)

K vecinos más cercanos (KNN) es un algoritmo de aprendizaje automático comúnmente utilizado para tareas de clasificación. Este emplea una medición de similitud, como la distancia Euclidiana, para comparar el dato de entrada con todos los demás datos de entrenamiento para así identificar los k vecinos más cercanos. Finalmente, para determinar la clase de la entrada, se realiza un voto mayoritario entre los datos más cercanos. En la metodología propuesta, se entrenó este modelo para poder clasificar a los individuos en base a las representaciones que retornaba la red Resnet-50.

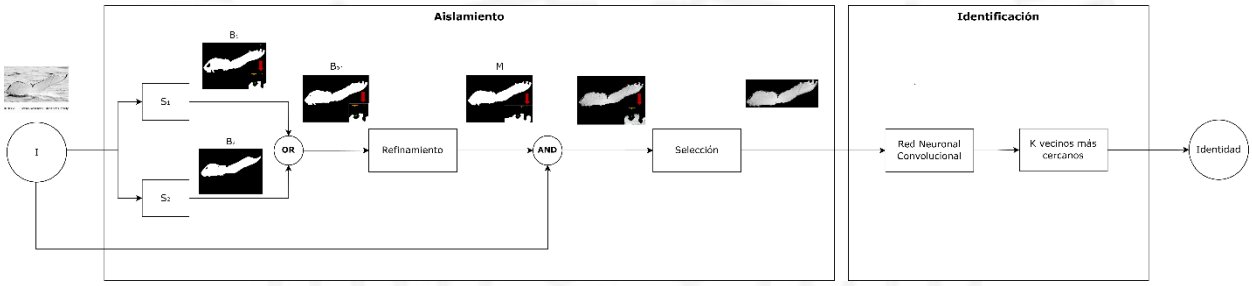
4. METODOLOGÍA

La foto-identificación constituye una herramienta importante para los biólogos marinos ya que ayuda a monitorear las poblaciones de diversas especies. Particularmente, los sistemas de foto-identificación de las ballenas jorobadas se sirven de técnicas de análisis de imágenes para encontrar patrones de pigmentación en las aletas caudales de las ballenas. Sin embargo, el utilizar imágenes tomadas en la naturaleza disminuye el rendimiento de identificación de estos sistemas ya que a menudo las imágenes contienen impurezas, como regiones con fondos de agua de mar y cielo. Intuitivamente, la segmentación de la aleta caudal de las ballenas jorobadas podría ayudar a mejorar los bajos rendimientos causados por tales artefactos; sin embargo, la evaluación de esta alternativa aún ha atraído la atención de la comunidad científica en la foto-identificación de ballenas. Por lo tanto, en esta investigación se propone evaluar el impacto de la segmentación de las aletas caudales de las ballenas en el proceso de foto-identificación, para lo cual se incorpora una etapa de aislamiento de la aleta caudal de las ballenas antes de la identificación propiamente dicha.

La etapa de aislamiento de la aleta caudal se basa en una combinación de técnicas de segmentación basadas en el aprendizaje profundo y procesamiento de imágenes, como se ilustra en la Fig. 4.1. Formalmente, dada una imagen de entrada, la etapa de segmentación primero produce un conjunto de potenciales máscaras de aleta caudal mediante la ejecución de dos técnicas de segmentación basadas en el aprendizaje profundo. A continuación, refina las máscaras para rellenar los posibles agujeros existentes mediante una técnica de procesamiento de imágenes. Por último, selecciona la máscara correspondiente a la aleta de la ballena. El proceso de foto-identificación se completa realizando la identificación sobre la imagen segmentada de la aleta caudal de la ballena. A continuación, se describe en detalle el diseño y el funcionamiento de las etapas de segmentación e identificación.

Figura 4.1

Metodología



4.1. Aislamiento de la aleta caudal

La etapa de aislamiento de la aleta caudal tiene como objetivo determinar la región correspondiente a la aleta caudal de la ballena a partir de una fotografía de su aleta capturada en la naturaleza. Para ello, esta etapa combina un conjunto de técnicas de segmentación semántica basadas en el aprendizaje profundo, un proceso de refinamiento y un proceso de selección, tal y como se representa en la Figura 4.1.

El proceso de aislamiento comienza con la segmentación de la aleta caudal de la ballena a partir de una imagen de entrada. Este proceso se lleva a cabo mediante técnicas de segmentación semántica (S) que producen imágenes binarias (B) en donde las agrupaciones de píxeles probablemente pertenezcan a la aleta caudal de la ballena. A continuación, un operador OR fusiona ambos resultados en una única imagen binaria (B_{or}) para mejorar la segmentación de la aleta caudal de la ballena. A continuación, una etapa de refinamiento de la región aplica un filtro morfológico para rellenar agujeros posiblemente existentes dentro de las regiones segmentadas (B_{or}). El resultado se comporta como una máscara (M) para las siguientes etapas en aislamiento de la aleta caudal. Posteriormente, un operador AND multiplica M con cada canal de la imagen de entrada para producir una imagen (M_y), cuyas regiones poseen una mayor probabilidad de pertenecer a la aleta caudal de la ballena. Finalmente, una etapa de selección de regiones escoge y recorta la región de la imagen (J) que está relacionada con la aleta caudal de la ballena, basándose en un criterio de ocupación de área.

A continuación, se describen las etapas más importantes del proceso de aislamiento de la aleta caudal de la ballena.

4.1.2. Segmentación semántica

La fase de segmentación semántica tiene el objetivo de identificar la región en la imagen de entrada (I), que pertenezca a la aleta caudal de la ballena jorobada. Castro & Ayma (2021) demostraron que las técnicas de segmentación basadas en aprendizaje profundo tienen un buen rendimiento en la diferenciación entre las aletas caudales y su entorno. Los mismos autores también mostraron que los enfoques de segmentación en su estudio podrían complementarse y dar lugar a un mejor rendimiento de segmentación de aleta caudal. De esta forma, es posible reunir

un conjunto de K técnicas de segmentación semántica basadas en el aprendizaje profundo para que trabajen juntas. Cada técnica de segmentación S_k, donde $k = 1, 2, 3 \ldots K$, produce una imagen binaria B_k; por ejemplo, en la Fig. 4.2 se puede observar las imágenes binarias B_1 y B_2 las cuales están conformadas por agrupaciones de píxeles C_n^i que señalan las posibles regiones de la aleta caudal de la ballena, donde n representa el número total de agrupaciones, como se detalla en el ejemplo de la Fig. 4.3. El proceso selección de la región de la imagen que corresponde a la aleta caudal de la ballena se presenta en la siguiente sección.

Figura 4.2

Imágenes binarias resultado de los algoritmos de segmentación S

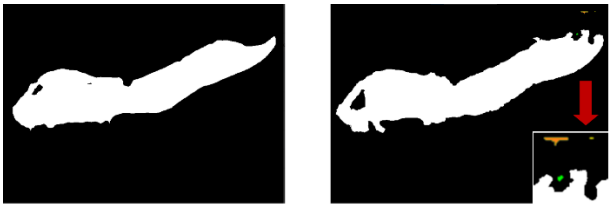
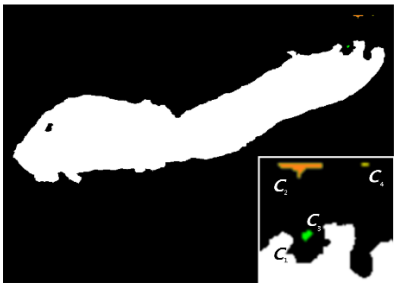


Figura 4.3

Agrupaciones de píxeles C_n^i, posibles regiones de aleta



En este trabajo, se reúnen dos técnicas de segmentación semántica basadas en el aprendizaje profundo para realizar la tarea de segmentación de la aleta caudal de la ballena. En concreto, se coloca a los algoritmos Red completamente convolucional (FCN) y Red de análisis de escenas en pirámide (PSP) a trabajar en dicha tarea. Esta elección se basa en el estudio de Castro & Ayma (2021), en donde las redes FCN y PSP mostraron rendimientos notables en la segmentación de aletas de ballena. Por lo tanto, se experimentó con la PSP construida sobre una ResNet101 y la variante FCN-8s de la FCN sobre la VGG-16. En ambos casos, se utilizaron las mismas configuraciones propuestas. A continuación, se describe la configuración utilizada durante el entrenamiento de los algoritmos.

4.1.3. Refinamiento

La etapa de refinamiento opera después de la fusión de los resultados de la segmentación de la aleta caudal de la ballena (B_k). Se espera que las técnicas de segmentación (S_k) proporcionen regiones de imagen fiables sobre la aleta caudal de la ballena; sin embargo, podrían producir máscaras de segmentación con agujeros debido a la naturaleza de la imagen de entrada. Dichas secciones podrían prevalecer después del proceso de fusión. Por lo tanto, esta etapa tiene como objetivo llenar los vacíos que quedan en el resultado de la fusión (B_{or}) para recoger la mayor cantidad de información posible de la aleta caudal de la ballena y evitar futuros fallos en la foto-identificación. Para esta tarea se utilizó la función binary_fill proporcionada por la biblioteca Scipy que rellena agujeros en máscaras binarias.

4.1.4. Selección

Una vez refinada la máscara binaria, esta se utilizará para generar la imagen segmentada realizando una operación lógica AND entre la máscara binaria resultante del proceso de selección y la imagen tridimensional original. A continuación, en la fase de selección, se seleccionará y recortará el área de la imagen segmentada que contiene información sobre el individuo. El objetivo de estas operaciones es perder la menor cantidad de información durante el redimensionamiento necesario para entrenar los algoritmos de aprendizaje profundo.

Para esta tarea, se empleó la función ConnectedComponentsWithStats2 proporcionada por la biblioteca OpenCV, la cual retorna el área de todos los componentes conectados en la imagen, permitiendo así seleccionar la región de la imagen con mayor área, asumiendo que esta se trata de la aleta caudal de la ballena jorobada. Este paso inicial fue necesario ya que, en algunas imágenes, existen agrupaciones de píxeles, a veces imperceptibles, fuera de la aleta caudal, lo que conlleva a un aislamiento errado de la aleta caudal. Posteriormente, se utilizó la función

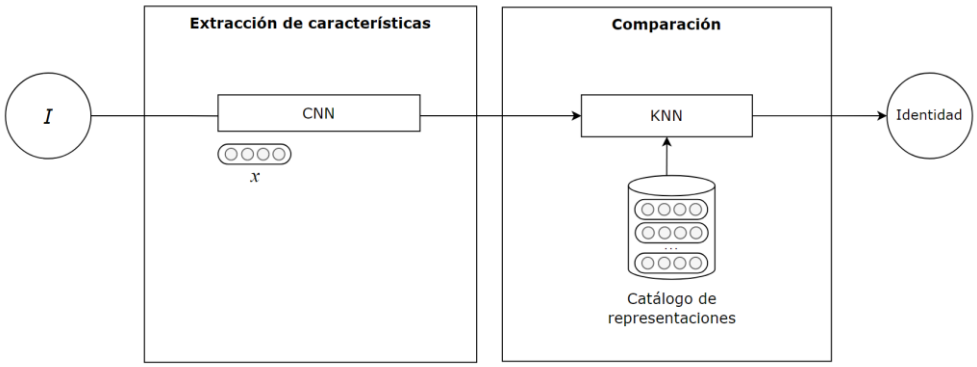
regionprops3 de la biblioteca scikit-image para recuperar el cuadro delimitador de la región seleccionada de la aleta caudal de la ballena jorobada. Finalmente, se recorta la imagen tomando como referencia este cuadro delimitador.

4.2. Identificación

En esta etapa se realiza la identificación de las ballenas jorobadas, considerando imágenes de sus aletas caudales que han pasado por el proceso de aislamiento explicado en la sección previa. El proceso de identificación se divide en dos fases: Extracción de características y comparación; los elementos de cada una de las fases se pueden observar en la Fig. 4.4. La primera fase emplea una arquitectura de CNN, entrenada previamente bajo el marco de aprendizaje de métricas de distancia y utilizando la función de Pérdida de tripletes, para retornar una representación x en base a las características distintivas de la imagen de entrada I. Posteriormente, la segunda fase se encargará de comparar las distancias entre esta representación y las representaciones de un catálogo previamente construido a partir de imágenes de ballenas jorobadas ya conocidas; para así, otorgar la identidad del más cercano. Si bien, esta comparación puede realizarse midiendo la distancia euclidiana entre las representaciones de imágenes, se optó por el uso del algoritmo K vecinos más cercanos (KNN) para facilitar esta tarea.

Figura 4.4

Fases de la etapa de identificación



A continuación, se presenta el detalle de las fases de Extracción de características y Comparación de la etapa de Identificación.

4.2.2. Extracción de características

Esta primera fase de identificación emplea una arquitectura de CNN entrenada con enfoque de aprendizaje de métricas de distancia para construir una representación x en base a las características extraídas de la imagen de entrada I. A continuación, se presentan los detalles del entrenamiento de la arquitectura de CNN.

Se seleccionó la arquitectura ResNet50, al igual que Bouma et al. (2018), ya que tiene un número reducido de capas que facilita su entrenamiento. Se añadió una capa de Dropout, Average Pooling y Dense, esta última con 256 neuronas y una operación de regularización L2 para devolver la representación x. Esta representación x constará de un arreglo de 256 valores numéricos que permitirá que la imagen sea comparada con otras. Fue necesario agregar estas capas adicionales para evitar el sobreajuste durante el entrenamiento, un problema muy común de Pérdida de tripletes (Schroff et al., 2015).

De forma empírica, se determinaron los ajustes de los parámetros que dieron los mejores resultados. Se inicializó la red ResNet50 con los pesos de su entrenamiento en el conjunto de datos ImageNet y se aplicaron operaciones de rotación y reflexión como aumento de datos. Las imágenes fueron redimensionadas a un tamaño uniforme de 444px de ancho y 244px de largo.

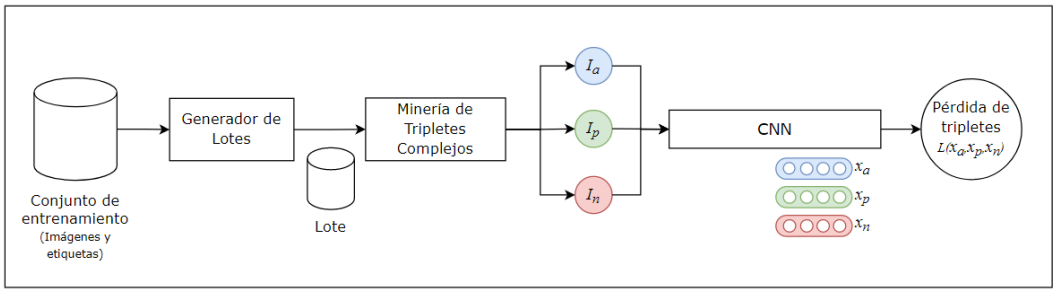
La red ResNet50 recibirá tres imágenes de entrada: un ejemplo ancla (I_a); un ejemplo positivo (I_p), una imagen de la misma clase que el ancla; y un ejemplo negativo (I_n) o imagen perteneciente a otra clase y retornará las representaciones correspondientes (x_a, x_p , x_n). Se utilizó la función de Pérdida de Tripletes con un margen de 1.0, un tamaño de lote de 64 imágenes, Adam como función de optimización con un ratio de aprendizaje de 0.0001 y se entrenó esta arquitectura con una configuración inicial de 100 épocas; sin embargo, se detuvo el entrenamiento cuando el valor de la pérdida dejó de disminuir durante tres épocas, con el objetivo de no caer en sobreajuste.

Finalmente, fue necesario construir nuestro propio generador de lotes que seleccionara estas 64 imágenes de forma aleatoria, pero garantizando que se pudieran ensamblar tripletes y que todas las imágenes disponibles para el entrenamiento se utilizarán el mismo número de veces. Para su construcción, se utilizaron las librerías Numpy y Pandas. Asimismo, se optó utilizar una estrategia de minería de tripletes complejos para conseguir mejores resultados.

En la Fig. 4.5 se puede observar una representación visual del proceso de entrenamiento de la fase de Extracción de características.

Figura 4.5

Entrenamiento de la fase de extracción de características



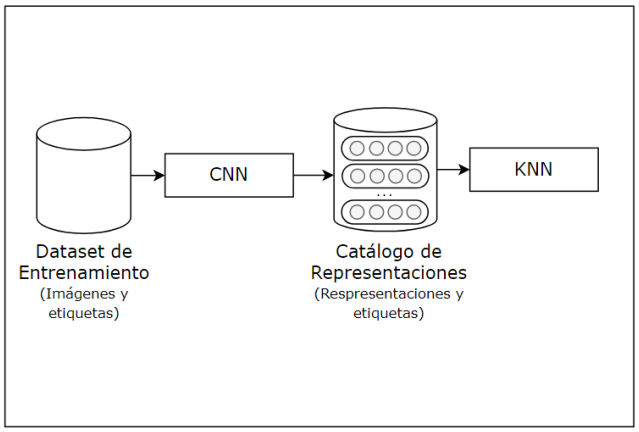
4.2.3. Comparación

Esta segunda fase de identificación emplea un modelo de K vecinos más cercanos (KNN) para poder predecir la identidad de la ballena jorobada basándose en las representaciones generadas previamente durante la fase de Extracción de características. A continuación, se presenta el proceso de entrenamiento de esta fase de identificación.

Para poder entrenar el modelo KNN, fue necesario generar las representaciones del subconjunto de imágenes de entrenamiento con la CNN previamente entrenada. Posteriormente, se utiliza este catálogo de representaciones, conformado por estas y sus respectivas etiquetas de identidad, como entrada al modelo KNN. Este utiliza la distancia euclidiana como métrica de distancia y, dada una nueva imagen, asignará la identidad perteneciente a la representación más cercana en el espacio. La Fig. 4.6 muestra una representación gráfica de este proceso de entrenamiento.

Figura 4.6

Entrenamiento de la fase de comparación



4.3. Conjunto de datos

Para evaluar el impacto del aislamiento de la aleta caudal de la ballena en el proceso de fotoidentificación, se utilizó un conjunto de datos de ballenas jorobadas ampliamente conocido y presentado en un concurso de Kaggle: *The Humpback Whale Identification Challenge* (Kaggle, 2019). Se optó por el uso de este conjunto de datos ya que consta de unas 25.000 imágenes de más de 4.000 individuos de ballena jorobada, lo cual lo convierte en la colección pública más amplia de fotografías de esta especie. Las imágenes de la aleta caudal de la ballena se tomaron en la naturaleza (en todo el mundo) utilizando diferentes dispositivos de captura de imágenes, lo que dio lugar a imágenes con diferentes tamaños (960x470 píxeles de media, con un máximo de 1050 y mínimo de 64 píxeles de alto y con un máximo de 1600 y mínimo de 30 píxeles de ancho), y propiedades fotométricas, como el contraste y el brillo. Además, la postura y el tamaño de las aletas de las ballenas en relación con las imágenes varían en el conjunto de datos. Además, dado que el conjunto de datos comprende imágenes de aletas de ballena jorobada tomadas en un entorno natural, puede haber oclusiones debidas a las olas, la espuma y las gotas de agua. La Fig. 4.7 muestra algunos ejemplos de imágenes del conjunto de datos.

Figura 4.7

Ejemplo del conjunto de datos

Nota. De Kaggle (2019)

5. EVALUACIÓN Y RESULTADOS

La presente investigación tiene como objetivo evaluar si la inclusión de un proceso de aislamiento previo a la identificación brinda mejores resultados en la foto-identificación de ballenas jorobadas. Con este propósito, se construyó un algoritmo de segmentación híbrido compuesto por las redes convolucionales FCN y PSPNet; además, se entrenó un algoritmo de identificación basado en aprendizaje de métricas de distancia con pérdida de tripletes. Finalmente, se evaluó el algoritmo de identificación en dos esquemas: sin utilizar el proceso de aislamiento y utilizando el proceso de aislamiento; de esta manera se busca corroborar dicha premisa.

En la presente investigación, se restringió el conjunto de datos presentado en un concurso de Kaggle: *The Humpback Whale Identification Challenge* (Kaggle, 2019) a aquellas imágenes que contaban con etiquetas. A continuación, se formaron los subconjuntos de entrenamiento, validación y prueba compuestos por el 70% (6369 imágenes de 3657 individuos), 10% (862 imágenes de 593 individuos) y 20% (1809 imágenes de 838 individuos), respectivamente. La partición se realizó de forma aleatoria, sin embargo, se utilizaron a los individuos con menos de 3 imágenes solamente en los conjuntos de datos de entrenamiento y validación. Además, los individuos con una imagen solo se emplearon como ejemplos negativos durante la construcción del triplete.

Para la evaluación de desempeño del algoritmo, se generaron las representaciones para las 1809 imágenes del conjunto de prueba y se compararon con las demás representaciones del conjunto de entrenamiento; de esta manera, se otorgó a la imagen la identidad del individuo que se encuentra a una menor distancia euclidiana. Esta evaluación se realizó tanto para el conjunto de imágenes sin aislamiento y con aislamiento. A continuación, se presentan los resultados.

Se calculó el *Mean Average Precision* (mAP) tanto para la primera como para las cinco imágenes más parecidas que retorna el algoritmo. En la Tabla 5.1 se puede observar que el modelo entrenado y evaluado sobre las imágenes con aislamiento alcanza un mayor mAP@1 de 0.66 y mAP@5 de 0.69; a comparación del que utiliza las imágenes sin aislamiento con puntajes de 0.51 y 0.53, respectivamente.

Tabla 5.1

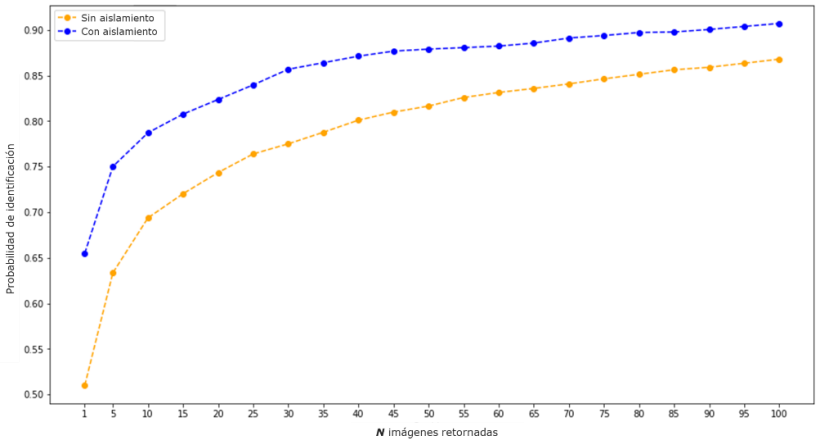
Mean Average Precision

Conjunto de imágenes	mAP@1	mAP@5
Sin aislamiento	0.5102	0.5310
Con aislamiento	0.6550	0.6906

Asimismo, se calculó el Cumulative Match Curve (CMC) para ambos esquemas, esta métrica indica la probabilidad de obtener la identidad correcta de una imagen nueva por cada *N* imágenes más parecidas retornadas por el algoritmo. En la Fig. 5.2 se pueden observar las curvas para el esquema, el algoritmo entrenado y evaluado sobre las imágenes con aislamiento, tiene una mayor probabilidad de reconocimiento en cada *N* imágenes retornadas a comparación del algoritmo entrenado y evaluado sobre las imágenes sin aislamiento, manteniéndose una diferencia promedio de 0.1. Por ejemplo, cuando el algoritmo retorna las 10 imágenes más parecidas, existe el 0.8 de probabilidad de que se haya retornado la identidad correcta en este conjunto si se utiliza el proceso de aislamiento; en cambio, cuando no se utiliza el proceso de aislamiento, solo se alcanza una probabilidad de 0.7.

Figura 5.2

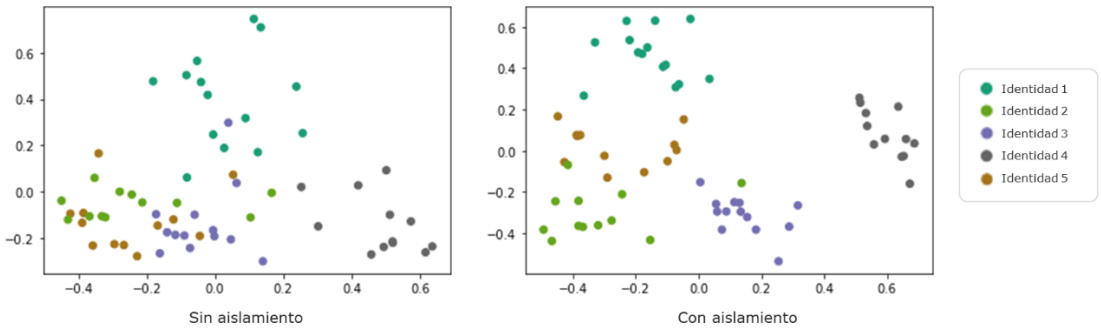
Curva CMC



En las Fig. 5.3 se muestra un ejemplo de la ubicación de las representaciones que genera cada esquema sobre imágenes de cinco individuos de nuestro conjunto de datos. Para ello, se redujeron las representaciones a dos dimensiones utilizando el algoritmo de PCA. Como se puede observar, las representaciones del modelo entrenado utilizando el proceso de aislamiento se encuentran mejor delimitadas. Esto influye en una mejor predicción de las identidades.

Figura 5.3

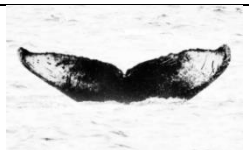
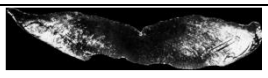
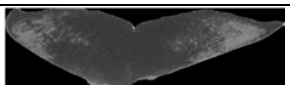
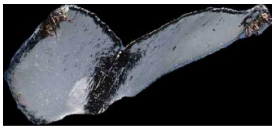
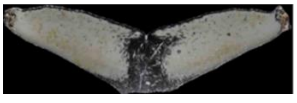
Representaciones de cinco individuos

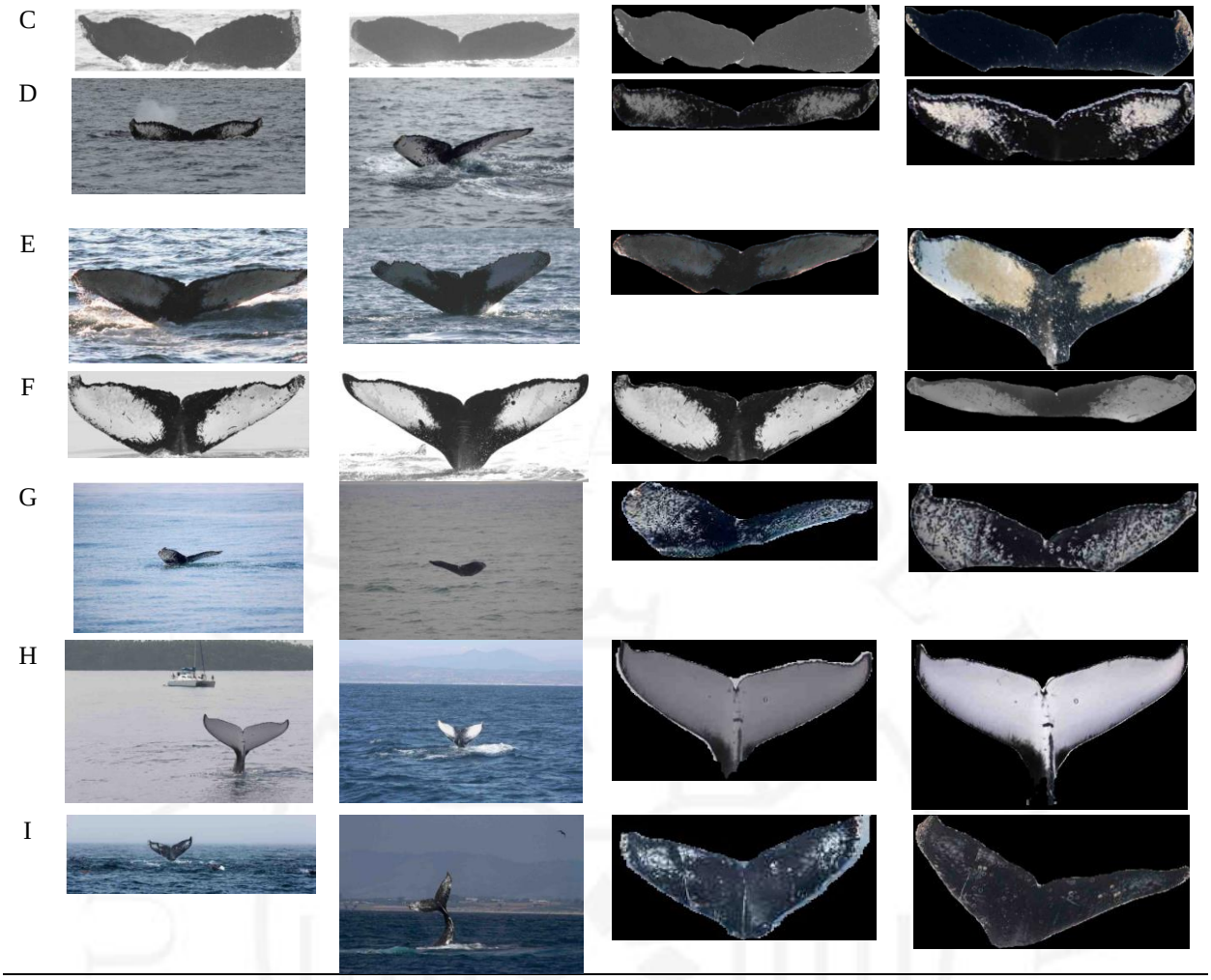


Por otro lado, en la Tabla 5.2 se puede observar un ejemplo de aquellas imágenes pertenecientes al conjunto de prueba que tienen una predicción incorrecta cuando se utiliza el esquema sin aislamiento, pero logran una predicción correcta usando aislamiento.

Tabla 5.2

Ejemplo de predicciones correctas usando aislamiento

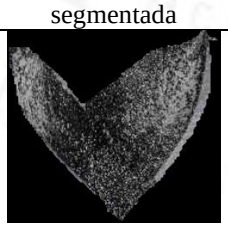
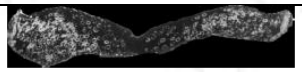
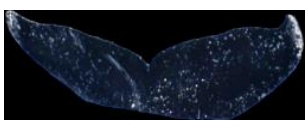
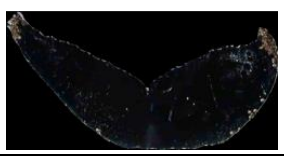
	Imagen no conocida sin aislamiento	Predicción incorrecta	Imagen no conocida con aislamiento	Predicción correcta
A				
B				



Finalmente, en la Tabla 5.3 se puede observar un ejemplo de aquellas imágenes pertenecientes al conjunto de prueba que, a pesar de seguir el proceso de aislamiento, no logran una predicción correcta.

Tabla 5.3

Ejemplo de predicciones incorrectas usando aislamiento

	Imagen no conocida segmentada	Predicción incorrecta
A		
B		
C		

6. DISCUSIÓN

El propósito de esta investigación es determinar si el aislamiento favorece los resultados de la foto-identificación automática de ballenas jorobadas. Para comprobar esta hipótesis, se construyó un algoritmo de segmentación híbrido basado en las arquitecturas FCN y PSPNet; además, se entrenó un algoritmo de identificación basado en aprendizaje de métricas de distancia con Pérdida de tripletes y ResNet50 como arquitectura base. Posteriormente, se comparó el rendimiento de este algoritmo de identificación sobre el conjunto de imágenes de prueba sin aislamiento y con aislamiento, mediante el cálculo del mAP y la curva CMC.

Bouma et al. (2018) también propone un modelo basado en pérdida de tripletes alcanzando un 90% de precisión; sin embargo, su propósito es identificar delfines comunes. Por otro lado, Schneider et al. (2020) utiliza el mismo conjunto de datos de *Humpback Whale Identification Challenge* (Kaggle, 2019) y realiza una comparación entre redes siamesas y pérdida de tripletes alcanzando un 0.75 de mAP; no obstante, no incluye ningún método de segmentación. La presente investigación dista del estado del arte ya que incorpora un proceso de segmentación durante la identificación de especímenes de ballena jorobada. Además, el propósito es analizar si este proceso de aislamiento previo mejora nuestros resultados de identificación.

Los resultados indican que el aislamiento permite tener mejores resultados en la identificación. Este preprocesamiento incrementa en 0.1 el Mean Average Precision y también logra una diferencia promedio de 0.1 en la curva CMC; llegando a un 0.75 de probabilidad de encontrar al individuo correcto cuando el algoritmo devuelve los 5 individuos más parecidos.

Estos resultados se deben a que, al utilizar la segmentación se descartan las partes irrelevantes de la imagen; así, el algoritmo de identificación se basa únicamente en las características del individuo, mas no en el fondo de la imagen. Por ejemplo, se puede observar en los ejemplos ubicados desde la fila A hasta la fila F de la Tabla 5.2, la predicción incorrecta tiene un gran parecido con la tonalidad y textura del fondo de la imagen original, retornando así un resultado erróneo que mejora cuando se utilizan las imágenes segmentadas.

Otro caso en donde la segmentación favorece a la identificación es cuando la aleta ocupa solo una pequeña porción de la imagen, filas G, H e I de la Tabla 5.2, aquí el proceso de segmentación, selección y recorte permite que se pierda una menor cantidad de información en la imagen al reajustar su tamaño para el entrenamiento y predicción de la red convolucional.

Finalmente, también existen imágenes en el conjunto de prueba que no logran una buena predicción a pesar de estar segmentadas y procesadas. Se pudo observar que, en ocasiones, artefactos como las gotas de agua se confunden como características adicionales al individuo; devolviendo así una predicción incorrecta (filas A, B y C, Tabla 5.3). Por esta razón, la incorporación de técnicas de eliminación de oclusiones puede ser tomado en cuenta para trabajos futuros.

7. CONCLUSIONES

El presente trabajo de investigación confirma que el aislamiento previo permite tener mejores resultados en la identificación de ballenas jorobadas. El modelo de segmentación híbrido basado en las redes PSPNet y FCN, en conjunto con el modelo de identificación basado en Pérdida de tripletes, logra un mAP de 0.66; a comparación de la utilización de este modelo de identificación con las imágenes sin segmentar, el cual alcanza un valor de 0.51. Además, en la curva CMC, la segmentación previa permite aumentar la probabilidad de identificación en 0.1 en promedio para cada ranking; alcanzando un 0.8 de probabilidad de encontrar al individuo correcto cuando las 10 imágenes más parecidas son retornadas. Este método propuesto, logra mejorar las predicciones descartando el fondo de la imagen, y enfocando al algoritmo de identificación en las características del individuo; además, favorece la identificación especialmente cuando la aleta ocupa una pequeña porción de la imagen original. Sin embargo, el método propuesto no logra buenos resultados cuando se presentan artefactos, como gotas de agua, que ocultan la aleta y se toman como características adicionales del individuo.

AGRADECIMIENTOS

Quiero expresar mi gratitud hacia la Universidad de Lima y su plana docente por brindarme los recursos académicos necesarios para llevar a cabo este trabajo de investigación. A mi asesor Oscar Efrain Ramos Ponce por su guía y consejos claves para la culminación del trabajo. Al profesor Victor Hugo Ayma Quirita por su dedicación y grandes enseñanzas durante la conceptualización y desarrollo de la investigación. A mi familia, por darme la oportunidad de formarme profesionalmente, por su constante aliento durante la elaboración de este trabajo y por siempre impulsarme a creer en mí.

REFERENCIAS

Ballance, L. (2018). Contributions of Photographs to Cetacean Science. *Aquatic Mammals*, 44(6), 668-682. https://doi.org/10.1578/am.44.6.2018.668

Barrientos, Y. (16 de julio de 2019). El avistamiento de ballenas jorobadas espera atraer a 45,000 turistas. El Correo. Recuperado de https://diariocorreo.pe/edicion/piura/el-avistamiento-de-ballenas-jorobadas-espera-atraer-45000- turistas-898801/

Barlow, J., Calambokidis, J., Falcone, E. A., Baker, C. S., Burdin, A. M., Clapham, P. J., ... Quinn, T. J. (2011). Humpback whale abundance in the North Pacific estimated by photographic capture-recapture with bias correction from simulation studies. *Marine Mammal Science*, 27(4), 793-818.

Bogucki, R., Cygan, M., Khan, C. B., Klimek, M., Milczek, J. K., y Mucha, M. (2018). Applying deep learning to right whale photo identification. In Conservation Biology (Vol. 33, Issue 3, pp. 676–684). Wiley.

Bouma, S., Pawley, M. D., Hupman, K., y Gilman, A. (2018). Individual Common Dolphin Identification Via Metric Embedding Learning. *2018 International Conference on Image and Vision Computing New Zealand (IVCNZ)*. Auckland, New Zealand.

Castro, A. y Ayma, V. H. (2021). *Humpback Whale's Flukes Segmentation Algorithms*. En: J. A. Lossio-Ventura, J. C. Valverde-Rebaza, E. Díaz. & H. Alatrista-Salas (Eds.) Information Management and Big Data: Seventh Annual International Conference, SIMBig 2020, Lima, Peru, October 1–3, 2020, Proceedings, Communications in Computer and Information Science (vol. 1410, pp. 291-303). Springer. https://doi.org/10.1007/978-3-030-76228-5_21

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., y Adam, H. (2018). Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. *Lecture Notes in Computer Science*, 833–851.

Elgendy, M. (2020). *Deep Learning for Vision Systems*. Manning.

Félix, F., Castro, C., Laake, J., Hasse, B., Scheidat, M. (2011). Abundance and survival estimates of the Southeastern Pacific humpback whale stock from 1991-2006 photo-identification surveys in Ecuador. *Journal of Cetecean Research and Management (Special Issue)*: 301-307

Ge, W., Huang, W., Dong, D., Scott, M.R. (2018). Deep Metric Learning with Hierarchical Triplet Loss. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds) Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science(), vol 11210. Springer, Cham. https://doi.org/10.1007/978-3-030-01231-1_17

Gestión (16 de julio de 2023). Temporada de avistamiento de ballenas en playas del norte generaría impacto de S/18 millones. El Correo. Recuperado de https://gestion.pe/peru/temporada-de-avistamiento-de-ballenas-en-playas-del-norte-generaria-impacto-de-s18-millones-noticia/

Gilman, A., Hupman, K., Stockin, K. A., y Pawley, M. D. (2016). Computer-assisted recognition of dolphin individuals using dorsal fin pigmentations. *2016 International Conference on Image and Vision Computing New Zealand (IVCNZ)*, 1-6

Gómez Blas, N., de Mingo López, L. F., Arteta Albert, A., y Martínez Llamas, J. (2020). Image Classification with Convolutional Neural Networks Using Gulf of Maine Humpback Whale Catalog. Electronics, 9(5), 731. MDPI AG.

He, K., Zhang, X., Ren, S., y Sun, J. (2015). Deep Residual Learning for Image Recognition (Versión 1). arXiv. https://doi.org/10.48550/ARXIV.1512.03385

Huang, G., Liu, Z., Van Der Maaten, L., y Weinberger, K. Q. (2017). Densely connected convolutional networks. *IEEE conference on computer vision and pattern recognition*, 4700-4708

Hsu, H. , Lee, Y., Ding, J., y Chang, R. (2018). Dolphin Recognition with Adaptive Hybrid Saliency Detection for Deep Learning Based on DenseNet Recognition. *2018 IEEE Asia Pacific Conference on Circuits and*

Systems (APCCAS). Chengdu, China.

Joly, A., Lombardo, J., Champ, J. y Saloma, A. (2016). Unsupervised Individual Whales Identification: Spot the Difference in the Ocean. *CLEF: Conference and Labs of the Evaluation Forum*, 2016.

Kaggle. (2019). Humpback Whale Identification Challenge. Recuperado de https://www.kaggle.com/c/humpback-whale-identification

Maglietta, R., Bruno, A., Reno, V., Dimauro, G., Stella, E., Fanizza, C., … Carlucci, R. (2018). The promise of machine learning in the Risso's dolphin Grampus griseus photo-identification. *2018 IEEE International Workshop on Metrology for the Sea; Learning to Measure Sea Health Parameters (MetroSea)*, 183-187.

Monnahan, C. C., Acevedo, J. , Noble Hendrix, A. , Gende, S. , Aguayo-Lobo, A. y Martinez, F. (2019), Population trends for humpback whales (Megaptera novaeangliae) foraging in the Francisco Coloane Coastal-Marine Protected Area, Magellan Strait, Chile. *Marine Mammal Science*, 35: 1212-1231.

Long, J., Shelhamer, E y Darrell, T. (2015). Fully Convolutional Networks for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4), 640-651.

Parkhi, O., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. BMVC 2015 - Proceedings of the British Machine Vision Conference 2015, 1–12.

Pollicelli, D., Coscarella, M., y Delrieux, C. (2020). RoI detection and segmentation algorithms for marine mammals photo-identification. *Ecological Informatics*, 56, 101038.

Ramos-Arredondo, R. I., Carvajal-Gámez, B. E., Gallegos-Funes, F. J., y Gendron-Laniel, D. (2014). Multi-Spatial Classifier for Blue Whale Images using Photo-Identification Method. *Research in Computing Science*, 82(2014), 31–40.

Reno, V., Dimauro, G., Labate, G., Stella, E., Fanizza, C., Capezzuto, F., … Maglietta, R. (2018). Exploiting species-distinctive visual cues towards the automated photo-identification of the Risso's dolphin Grampus griseus. *2018 IEEE International Workshop on Metrology for the Sea; Learning to Measure Sea Health Parameters (MetroSea).*

Ruiz Effio, M. (2016). Ballenas en el norte del Perú. Lima: Fondo Editorial Universidad Científica del Sur.

Safina, C. (6 de junio del 2016). The unseen significance of whales. Recuperado de https://blog.nationalgeographic.org/2016/06/01/the-unseen-significance-of-whales/.

Shanmugamani, R. (2018). *Deep learning for computer vision: expert techniques to train advanced neural networks using TensorFlow and Keras.* Birmingham, UK: Packt Publishing.

Shi, H., Yang, Y., Zhu, X., Liao, S., Lei, Z., Zheng, W., y Li, S. Z. (2016). Embedding Deep Metric for Person Re-identification: A Study Against Large Variations. In Computer Vision – ECCV 2016 (pp. 732–748). Springer International Publishing. https://doi.org/10.1007/978-3-319-46448-0_44

Schneider, S., Taylor, G., y Kremer, S. (2020). Similarity Learning Networks for Animal Individual Re-Identification. *The IEEE Winter Conference on Applications of Computer Vision (WACV)*. 44-52

Schroff, F., Kalenichenko, D., y Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 815–823. https://doi.org/10.1109/CVPR.2015.7298682

Song, H. O., Xiang, Y., Jegelka, S., y Savarese, S. (2015). Deep Metric Learning via Lifted Structured Feature Embedding. arXiv. https://doi.org/10.48550/ARXIV.1511.06452

Tao, R., Gavves, E., y Smeulders, A. W. M. (2016). Siamese Instance Search for Tracking. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. https://doi.org/10.1109/cvpr.2016.158

Tensorflow (s.f.). TensorFlow Addons Losses: TripletSemiHardLoss https://www.tensorflow.org/addons/tutorials/losses_triplet

Thompson, J. W., Zero, V. H., Schwacke, L. H., Speakman, T. R., Quigley, B. M., Morey, J. S. M., y McDonald, T. L. (2019). finFindR: Computer-assisted Recognition and Identification of Bottlenose Dolphin Photos in R. Cold Spring Harbor Laboratory. https://doi.org/10.1101/825661

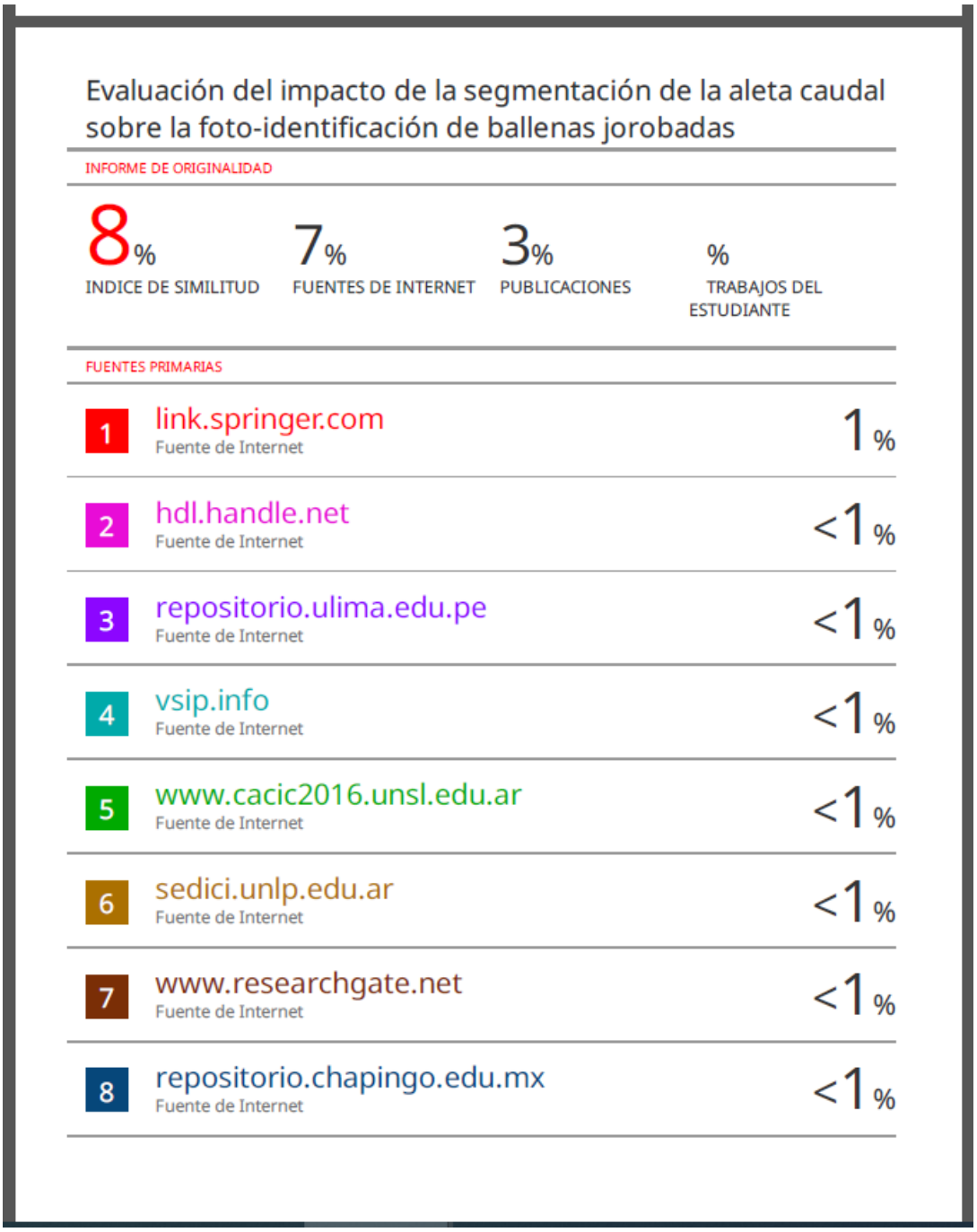
Titova, O. V., Filatova, O. A., Fedutin, I. D., Ovsyanikova, E. N., Okabe, H., Kobayashi, N., ... Hoyt, E. (2018). Photo-identification matches of humpback whales (Megaptera novaeangliae) from feeding areas in Russian Far East seas and breeding grounds in the North Pacific. *Marine Mammal Science*, 34(1), 100-112.

Ustinova, E., y Lempitsky, V. (2016). Learning Deep Embeddings with Histogram. arXiv. https://doi.org/10.48550/ARXIV.1611.00822

Weideman, H., Jablons, Z. Holmberg, J., Flynn, K., Calambokidis, J., Tyson, R., … Stewart, C. (2017). Integral Curvature Representation and Matching Algorithms for Identification of Dolphins and Whales. *2017 IEEE International Conference on Computer Vision Workshop (ICCVW)*. Venice, Italy.

Wohlhart, P., & Lepetit, V. (2015). Learning Descriptors for Object Recognition and 3D Pose Estimation. arXiv. https://doi.org/10.48550/ARXIV.1502.05908

Wurm, M., Stark, T., Zhu, X. X., Weigand, M., y Taubenböck, H. (2019). Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks. En ISPRS Journal of Photogrammetry and Remote Sensing (Vol. 150, pp. 59-69). Elsevier BV. https://doi.org/10.1016/j.isprsjprs.2019.02.006



Evaluación del impacto de la segmentación de la aleta caudal sobre la foto-identificación de ballenas jorobadas

INFORME DE ORIGINALIDAD

8%	7%	3%	%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	link.springer.com Fuente de Internet	1%
2	hdl.handle.net Fuente de Internet	<1%
3	repositorio.ulima.edu.pe Fuente de Internet	<1%
4	vsip.info Fuente de Internet	<1%
5	www.cacic2016.unsl.edu.ar Fuente de Internet	<1%
6	sedici.unlp.edu.ar Fuente de Internet	<1%
7	www.researchgate.net Fuente de Internet	<1%
8	repositorio.chapingo.edu.mx Fuente de Internet	<1%