Revista Iberoamericana de Automática e Informática Industrial

Revista Iberoamericana de Automática e Informática Industrial 16 (2019) 358-368

www.revista-riai.org

Fusión temprana de descriptores extraídos de mapas de prominencia multi-nivel para clasificar imágenes

E. Fidalgo[a,c,*], E. Alegre[a,c], L. Fernández-Robles[b,c], V. González-Castro[a,c]

[a] *Departamento de Ingeniería Eléctrica, y de Sistemas y Automática. Universidad de León, Spain*
[b] *Departamento de Ingenierías Mecánica, Informática y Aeroespacial. Universidad de León, Spain*
[c] *Researcher at INCIBE (Spanish National Institute of Cybersecurity), León, Spain*

Resumen

En este artículo proponemos un método que permite mejorar la clasificación de imágenes. Considerando los mapas de prominencia como mapas topográficos y filtrando las características del fondo de la imagen, se mejora la codificación que realiza el modelo de Bag of Visual Words (BoVW). Primero, evaluamos seis algoritmos para la generación de mapas de prominencia, seleccionando GBVS y SIM porque son los que retienen la mayor parte de la información del objeto. Después, eliminamos los descriptores SIFT extraídos pertenecientes al fondo mediante el filtrado de características en base a imágenes binarias obtenidas a diversos niveles o capas de dichos mapas de prominencia. Por último, evaluamos la fusión temprana de los descriptores SIFT filtrados en cinco conjuntos de datos diferentes. Los resultados obtenidos indican que el método propuesto mejora siempre al método de referencia cuando se combinan las dos primeras capas de GBVS o de SIM y el dataset contiene imágenes con un único objeto.

Palabras Clave:
Visión por computador, Algoritmos de detección, Aprendizaje máquina, Procesamiento de imágenes, Codificación, Clasificadores

Early Fusion of Multi-level Saliency Descriptors for Image Classification

Abstract

In this paper, we propose a method that improves the classification of images. Considering saliency maps as if they were topographic maps and filtering the characteristics of the image's background, the Bag of Visual Words (BoVW) coding is improved. First, we evaluated six known algorithms to generate saliency maps and we selected GBVS and SIM because they are the ones that retain most of the information of the object. Next, we eliminated the extracted SIFT descriptors belonging to the background by filtering features based on binary images obtained at various levels of the selected saliency maps. We filtered the descriptors by obtaining layers at various levels of the saliency maps, and we evaluated the early fusion of the SIFT descriptors contained in these layers into five different datasets. The results obtained indicate that the proposed method always improves the reference method when combining the first two layers of GBVS or SIM and the dataset contains images with a single object.

Keywords:
Computer Vision, Detection algorithms, Machine Learning, Image processing, Encoding, Classifiers

1. Introducción

Uno de los pasos críticos en la clasificación de imágenes es la extracción de características, o descripción de las imágenes. SIFT (del inglés *Scale Invariant Feature Transform*) (Lowe, 2004) es un descriptor que continúa, aún hoy, obteniendo resultados del estado del arte (González-Castro et al., 2017; He et al., 2018; Al-khafaji et al., 2018). Muchos trabajos recientes usan SIFT como un estándar para comparar con los resultados de los métodos que proponen (Trzcinski et al., 2015; Fidalgo et al., 2016, 2017).

El modelo de la Bolsa de Palabras Visuales (BoVW, del inglés, *Bag of Visual Words*) (Csurka et al., 2004) representa cada imagen como un vector de características usando un

* Autor para correspondencia: efidf@unileon.es
To cite this article: E. Fidalgo, E. Alegre, L. Fernández-Robles, V. González-Castro. 2019. Sliced weightened schema for improved image classification. Revista Iberoamericana de Automática e Informática Industrial 16, 358-368. https://doi.org/10.4995/riai.2018.10640

diccionario visual creado previamente a partir de los descriptores mediante un proceso de agrupamiento, o *clustering*. Cada componente de este vector de características representa la frecuencia con la que cada palabra visual del diccionario aparece en la imagen (Biagio et al., 2014). BoVW todavía muestra un rendimiento notable en procesos de clasificación y de recuperación de imágenes (*image retrieval*). Los investigadores todavía utilizan a menudo este método para realizar eficazmente la recuperación de imágenes representadas mediante características globales en imágenes duplicadas (Chatzichristofis et al., 2013), cálculo de diccionarios visuales mejorados (Zheng et al., 2014) o esquemas de ponderación (Zheng et al., 2013), y también añaden información espacial a las palabras visuales (Chen et al., 2014b). Para una revisión más extensiva, Chen et al. (2014a) presentan las tendencias actuales en recuperación de imágenes usando el modelo BoVW.

Si se usan los descriptores de la imagen completa directamente para crear el diccionario visual, se codifican vectores de características que contendrán información tanto de los objectos de interés en primer plano como del fondo, lo que puede resultar en clasificaciones subóptimas (Borji and Itti, 2013).

Entre las diferentes estrategias para seleccionar automáticamente los objetos de interés, en este artículo nos hemos centrado en las técnicas de prominencia visual (*visual saliency*), es decir, la medida de los estímulos a bajo nivel que captan la atención humana en las primeras etapas del procesamiento visual (Itti et al., 1998). Existen otras estrategias para localizar objetos en una imagen, como las "propuestas de regiones" (*region proposals*) (Saikia et al., 2017; Sepúlveda et al., 2017; Chaves et al., 2018; Saikia et al., 2018; García-Olalla et al., 2018), pero su revisión queda fuera del ámbito de este artículo. La representación y medida de la prominencia visual es un tema que ha sido atractivo para la comunidad de Visión por Computador (Borji and Itti, 2013). Lahouli et al. (2018) presentaron un método de detección de personas cuya primera etapa se basa en la obtención de un mapa de prominencia visual. Por su parte, Fang et al. (2017) propusieron un método para el aprendizaje de la prominencia visual en imágenes estereoscópicas, que toma en cuenta la información de la profundidad. Jian et al. (2018) desarrollaron un sistema para generar mapas de prominencia visual mediante segmentación basada en superpíxeles. Murray et al. (2013) intentaron introducir conocimiento a priori en el proceso de prominencia adaptando el modelo de inducción de color de bajo nivel antes de predecir la prominencia. Hou et al. (2012) aislaron el soporte del fondo en el domino transformado de señales mezcladas en una imagen I y posteriormente lo transformaron de nuevo al dominio espacial calculado la imagen reconstruida. Fidalgo et al. (2018) demostraron que calcular los descriptores únicamente de los objetos de interés, sin tener en cuenta el fondo de la imagen, produce información más discriminatoria, lo que mejora la tasa de acierto en tareas de clasificación de imágenes.

Un mapa de prominencia se puede considerar como un mapa topográfico que registra el nivel de prioridad de atención visual. Así, encontramos que, dependiendo de la *altura* de un corte imaginario en dicho mapa topográfico, se podría obtener una vista diferente, o nivel de información de la imagen analizada tras la superposición de los planos binarios correspondientes a cada uno de los diferentes ocho bits de la representación de una imagen en escala de grises. Nos referiremos a estos *planos* a una cierta *altura* en el mapa topográfico, es decir, en el mapa de prominencia, como *capas*. La Figura 1 ilustra la propuesta realizada en este artículo.

En este trabajo vamos a evaluar la influencia de la información contenida en un mapa de prominencia en diferentes planos, o capas binarias, de manera individual o combinadas con el modelo BoVW. El uso de múltiples capas en una imagen o en un mapa de prominencia no es una idea nueva, aunque hasta donde sabemos, sí lo es su uso combinado con el modelo BoVW para la clasificación de imágenes. Yan et al. (2013) usaron múltiples capas de la imagen original. En primer lugar aplicaron *Watershed* (Digabel and Lantuéjoul, 1978; Beucher and Lantuejoul, 1979; Gonzalez and Woods, 2002) para obtener una imagen en color sobre-segmentada y, posteriormente, fusionaron regiones en tres escalas diferentes, una por capa, por lo que, para cada escala, obtuvieron una imagen en color con regiones fusionadas. Una vez obtenidas estas capas de color calcularon sus prominencias basándose en el contraste local y localizaciones heurísticas. Finalmente, las combinaron para obtener el mapa de prominencia utilizado. Margolin et al. (2013) presentaron un método basado en cortes a diferentes niveles de prominencia modificando el parámetro H que controla la selección de los píxeles altamente distintivos (HDP, *del inglés Highly Distinctive Pixels*). En dicho trabajo, explicaron que las diferentes capas contienen varios niveles de información, pero no extrajeron las características de dichas capas ni analizaron su contribución a un proceso de clasificación de imágenes. En nuestro último trabajo (Fidalgo et al., 2018), binarizamos el mapa de prominencia obtenido mediante el método de Hou et al. (2012), para crear una máscara binaria que superpusimos con la imagen original para filtrar las áreas de atención. Sin embargo, al binarizar el mapa de prominencia no tuvimos en cuenta la información de las otras capas que contenían los diferentes planos binarios.

En este trabajo, después de revisar y seleccionar seis métodos que generan diferentes mapas de prominencia (Hou et al., 2012; Harel et al., 2007; Itti et al., 1998; Murray et al., 2013; Zhang et al., 2013; Vikram et al., 2012), aplicamos dichos mapas a 40 imágenes de dos subconjuntos de la base de datos ImageNet (Russakovsky et al., 2015) y, tras un proceso de selección cualitativo, elegimos dos de estos algoritmos para generar los mapas que usamos con los distintos esquemas de codificación empleados para clasificar imágenes. Dicho proceso de selección se basó en un sistema de votación sobre esas 40 imágenes, donde se asignó un mayor número de votos al mapa de prominencia que mejor seleccionaba los objetos de interés en la mayoría de las imágenes. Tras el proceso de selección, evaluamos la tasa de acierto obtenida al clasificar imágenes basándonos en cada uno de estos dos mapas de atención. En primer lugar, para cada uno de ellos calculamos el mapa de prominencia de la imagen. A continuación, binarizamos dichos mapas mediante el método de Otsu (Otsu, 1979), obteniendo la capa correspondiente a un nivel de prominencia determinado. La máscara binaria obtenida resalta el primer plano de la imagen, del cual se extraen los descriptores que se usan en el modelo BoVW. En este trabajo se extrae información de cuatro niveles de prominencia diferentes, seleccionados como un porcentaje del nivel de prominencia obtenido mediante la binarización de Otsu. Se ha realizado este proceso con cinco conjuntos de imágenes públicos, dos de los

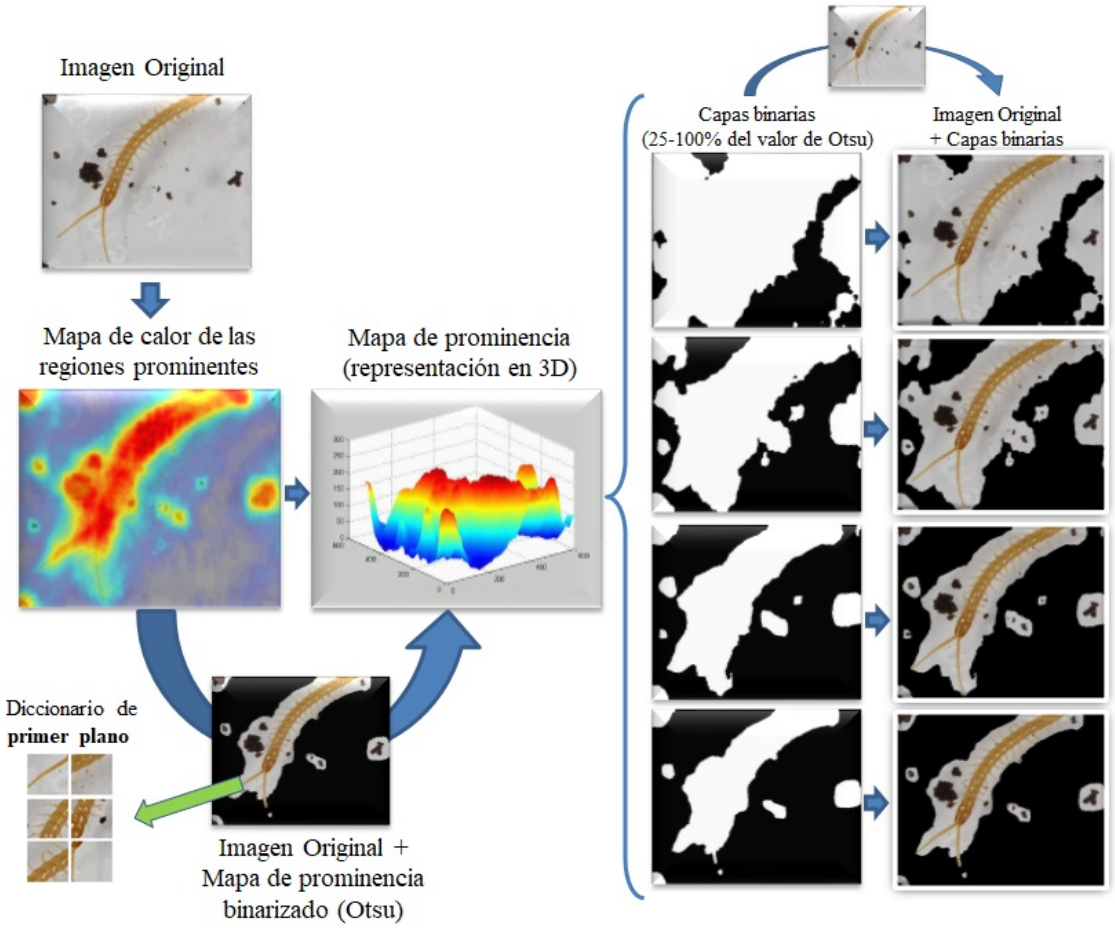

Figura 1: Esquema gráfico de la selección de características del mapa de prominencia por capas

cuales son subconjuntos de ImageNet. Tras evaluar la cantidad de la información contenida en cada capa, combinamos la información de cada capa antes de la construcción del diccionario, usando un esquema de fusión temprana. De esta manera, evaluamos cuánto contribuye la información de cada capa al modelo final entrenado. Finalmente, combinamos la información de los dos mapas de prominencia utilizados en forma de un descriptor para demostrar que la combinación de capas o niveles de información de diferentes mapas de prominencia se puede utilizar para mejorar los resultados de referencia.

En resumen, las principales contribuciones de este trabajo con respecto al estado del arte se pueden enumerar como sigue:

- En el contexto de la clasificación de imágenes, se propone el uso de mapas de prominencia visual combinado con su división en diferentes niveles para mejorar la descripción de los objetos de interés.

- En particular, se usa el modelo BoVW, con SIFT como descriptor de bajo nivel, para describir las imágenes. Una de las principales novedades de este trabajo es que, por primera vez, los diccionarios del modelo BoVW se codifican únicamente con la información de la imagen que está contenida en los mapas de prominencia visual a diferentes niveles. Esto permite que los diccionarios codificados de esta manera sean más discriminativos y, por ello, se obtengan mejores resultados en la clasificación.

- Además, con el objetivo de que los descriptores ubicados sobre los objetos resaltados por el mapa de prominencia tengan un mayor peso en el cálculo del BoVW, se obtienen diversas capas cortando el mapa de prominencia a diferentes niveles, y luego se realiza la fusión temprana de los descriptores obtenidos. De esta manera, se obtiene una tasa de acierto superior a la que proporcionan los descriptores SIFT de toda la imagen para codificar los diccionarios.

El resto del artículo se organiza de la siguiente manera: en la Sección 2 se describen los seis mapas de prominencia estudiados, así como el proceso realizado para seleccionar dos de ellos que fueron los utilizados posteriormente en el método presentado. En la Sección 3 se explican cómo se obtiene la información de cada nivel de información o capa, así como las diferentes fusiones de descriptores evaluadas. A continuación, en la Sección 5 se detallan los experimentos llevados a cabo y se realiza una discusión de los resultados. Finalmente, las conclusiones y perspectivas futuras se presentan en la Sección 6.

2. Estudio y selección de mapas de prominencia

En esta Sección, se presenta una breve descripción de tres mapas de prominencia inspirados biológicamente (ver Secciones 2.1, 2.2, 2.3) y otros inspirados en aprendizaje automático

(ver Secciones 2.4, 2.5, 2.6), además de una comparación gráfica entre ellos. A continuación se muestra la selección de los dos mapas de prominencia utilizados en nuestros experimentos.

2.1. Mapa de prominencia de Itti

Principalmente, hay dos tipos de aproximaciones para modelar la prominencia en una imagen (Borji and Itti, 2013), unas basada en modelos inspirados biológicamente y otras en aprendizaje automático. El mapa de prominencia propuesto por Itti et al. (1998) se basa en los primeros, en concreto en la arquitectura del sistema visual de los primeros primates. Se calculan 42 mapas de características, seis para intensidad, 12 para color y 24 para ubicación usando una pirámide Gausiana (Greenspan et al., 1994) creada a partir de la imagen original. Después, todos ellos se combinan a través de un operador de normalización, que selecciona mapas con un pequeño número de altos picos de actividad y no tiene en cuenta mapas con numerosas respuestas de picos comparables. De forma resumida, el proceso que se sigue es el siguiente. Primero se normalizan a un rango fijo todos los valores de los mapas. Luego se calcula el valor promedio de todos los valores máximos globales del mapa y, finalmente, el mapa completo se multiplica por una determinada constante. A partir de los mapas normalizados anteriores, se construyen tres mapas de visibilidad, uno para la intensidad, otro para el color y el último para la ubicación, a través de la adición a gran escala, es decir, la reducción de cada mapa a escala 4 y una adición punto por punto. Finalmente, los tres mapas de visibilidad se suman y promedian para obtener el mapa de prominencia final, al que haremos referencia en este trabajo como *Itti*.

2.2. Prominencia visual basada en grafos

Harel et al. (2007) propusieron un modelo de prominencia que consiste en crear mapas de activación en canales de características, luego normalizarlos para resaltar áreas esenciales y finalmente combinarlas con otros mapas. Los autores siguen las etapas clásicas de los modelos de prominencia visual: primero extraen vectores de características, luego forman un mapa de activación con estos vectores y finalmente lo normalizan. El núcleo de este método es la forma en que calculan el mapa de activación, a través de un enfoque Markoviano, midiendo la diferencia entre dos puntos. En este punto, los autores consideran su aproximación "órgánica", debido a que, biológicamente, los nodos del grafo (neuronas) se conectan a través de una red (cortex visual) y se comunican entre ellos (activaciones sinápticas). Al final, se resaltan los lugares importantes donde la imagen tiene información de acuerdo con la fijación humana. Nos referiremos a este mapa de prominencias como *GBVS*.

2.3. Prominencia a través de mecanismos de inducción

Los principales desafíos de los mapas de prominencia inspirados biológicamente son: generar mapas óptimos de características antes de estimar los mapas de prominencia (Kienzle et al., 2007), combinar la información de prominencia de los mapas de características (Zhao and Koch, 2012) y seleccionar los parámetros del modelo en todo el proceso (Pinto et al., 2009), como las características o coeficientes de filtros para las funciones de activación y normalización. Murray et al. (2013) propusieron un mecanismo de inducción de prominencia (SIM) basado en un modelo espacial-cromático de bajo nivel. Esta propuesta intenta abordar los problemas mencionados anteriormente en los mapas de prominencia inspirados biológicamente adaptando el modelo de inducción de color de bajo nivel antes de predecir la prominencia, como un conocimiento previo insertado en el proceso que también ayuda con el ajuste de parámetros. SIM utiliza gránulos geométricos (Mallat, 2009) para obtener una representación de la imagen final que resalte las características más prominentes mientras suprime las menos destacadas. Nos referiremos a este mapa de prominencia como *SIM*.

2.4. Prominencia circundante de centros aleatorios

Vikram et al. (2012) propusieron calcular el mapa de prominencia ayudados por las prominencias locales que aparecían en regiones rectangulares de interés. Al igual que el algoritmo SDSP (que se explicará en la Subsección 2.6), se convierte la imagen inicial en el espacio de color Lab y se trabaja individualmente en cada canal, realizando una división aleatoria de subventanas. El valor de prominencia de cada píxel se define como la diferencia de la intensidad del píxel y la intensidad media de la ventana secundaria que lo contiene. Por lo tanto, el mapa de prominencia final tiene la misma dimensión que la imagen original. Las coordenadas de las subventanas se generan a través de una función de distribución de probabilidad uniforme discreta. Durante nuestro trabajo, denominaremos a este mapa de prominencia *RCSS*, y en él, el único parámetro ajustable es el número de ventanas secundarias que se calcularán por canal.

2.5. Algoritmo de prominencia basado en la firma de la imagen

En uno de nuestros trabajos previos (Fidalgo et al., 2018) utilizamos el algoritmo de Hou et al. (2012) para demostrar cómo el uso de diferentes factores de desenfoque del algoritmo afectaba la precisión obtenida en la clasificación de imágenes utilizando BoVW. Este método considera una imagen como la suma de dos señales: de primer plano y de fondo y se basa en la firma de la imagen, I_{sig}, definida de la siguiente manera:

$$I_{sig} = sign(DCT(I)), \quad (1)$$

donde *sign* es la función de signo y *DCT* es la transformada discreta del coseno. Nos referiremos a este mapa de prominencia como *Image Signature*.

2.6. Detección de prominencia basada en combinar información previa

Las personas tendemos a centrar nuestra visión en el centro de una imagen, y nos resultan más atractivos los colores cálidos que los fríos. En función de cómo el sistema visual humano detecta las regiones de prominencia, Zhang et al. (2013) propusieron una detección de prominencia que resulta de la combinación de información previa sobre la frecuencia, color y la ubicación. Partiendo del trabajo de Toet et al. (2009), calcularon la frecuencia con un filtro de Gabor logarítmico (Field, 1987) en lugar de con la diferencia de Gaussianos como hicieron Toet et al. (2009). Existen estudios (Shen and Wu, 2012) que demuestran que los colores cálidos son más llamativos para el sistema visual humano que los fríos. Basándose en eso,

Zhang et al. (2013) calcularon la información de color convirtiendo la imagen del espacio RGB al Lab, escalando los componentes entre [0,1] y asignando a 1 la mayoría de las regiones de prominencia, y 0 el resto de ellas. Dado que se ha demostrado que los objetos cercanos al centro de una imagen son más atractivos para las personas (Tilke et al., 2009), Zhang et al. (2013) propusieron que las ubicaciones cercanas al centro serían más prominentes que el resto de la imagen, por lo que modelan la ubicación con un mapa Gaussiano. El mapa de prominencia final, al que a partir de ahora nos referiremos como *SDSP*, es el producto de los tres antecedentes ya explicados.

2.7. *Evaluación y selección de los mapas de prominencia*

En la Figura 2 y en la Figura 3 puede verse la representación visual del mapa de prominencias explicado en las subsecciones anteriores sobre algunas muestras de imágenes pertenecientes a dos subconjuntos extraídos del conjunto de datos ImageNet. Los hemos denominado ImageNet-7Birds e ImageNet-7Arthropods, presentando más detalles sobre estos conjuntos en la subsección 4.1.

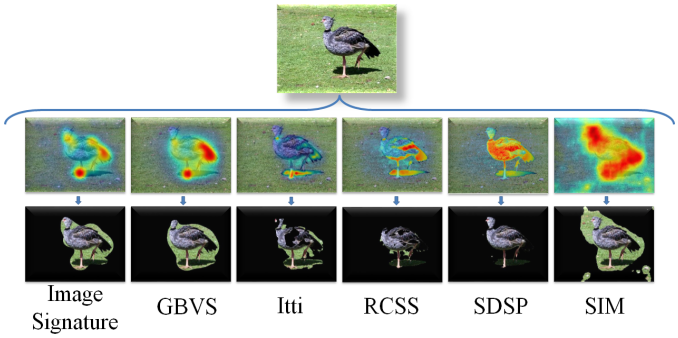

Figura 2: Ejemplo de imagen de ImageNet-7Birds (Fila 1), el mapa de calor asociado al mapa de prominencia aplicado (Fila 2) y el mapa de prominencia tras binarizarlo y superponerlo con la imagen original (Fila 3)

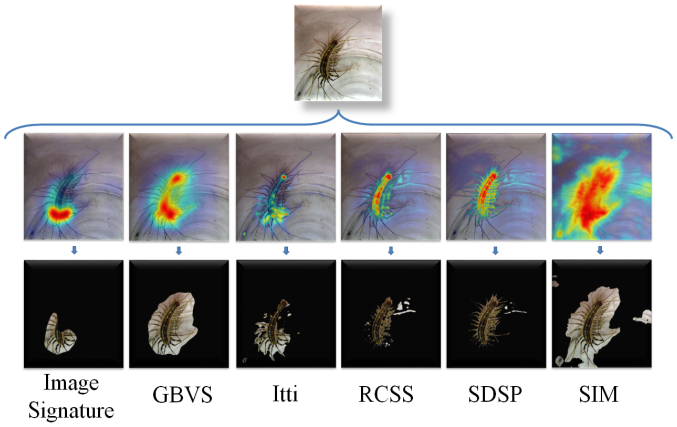

Figura 3: Imagen de ejemplo de ImageNet-7Arthropods después de aplicar los seis mapas de prominencia estudiados

Nuestro objetivo es la evaluación de la influencia de la información contenida a diferentes niveles de un mapa de prominencia. Utilizando dichos mapas como filtros de características, vamos a tratar de mejorar la codificación de características en el modelo BoVW y obtener resultados superiores a los que se obtendrían sin utilizar dicha información extraída de diferentes capas de mapas.

Con la intención de establecer cuáles eran los mapas más apropiados para la anterior tarea, seleccionamos únicamente dos de los seis mapas de prominencia presentados previamente. Para ello, aplicamos los seis algoritmos anteriores a todas las imágenes de los subconjuntos ImageNet-7Birds y ImageNet-7Arthropods, y visualizamos el resultado obtenido al superponer a la imagen original, tanto el mapa de prominencia, en la primera fila, como una imagen binaria obtenida mediante el método de Otsu, en la segunda de las Figuras 2 y 3. Elegimos al azar 40 imágenes de cada uno de dichos conjunto de datos para realizar una inspección visual detallada y aplicamos un sistema de votación que nos permitió seleccionar los dos métodos más apropiados para este problema.

En este trabajo nos hemos centrado en el problema de la clasificación de imágenes fino (en inglés, “fine-grained image classification”). Por este motivo, se decidió realizar una selección cualitativa de mapas de prominencia, representada por una decisión binaria para cada imagen sobre la presencia del objeto de interés en el mapa de prominencia, es decir la asignación de la etiqueta “válido” (1) o “no válido” (0), según incluyera o no, respectivamente, el objeto de interés. El motivo de realizar este sistema de votación, en lugar de un sistema de puntuaciones por rango (por ejemplo), está basado en nuestro interés de determinar qué capa o capas de un mapa de atención, número y posibles combinaciones aportan más información a un proceso de clasificación de imágenes y permiten superar unos resultados de referencia obtenidos sin mapas de promiencia. Evaluar diferentes niveles de calidad de resultados utilizando puntuaciones en un rango queda fuera del alcance de este estudio.

En las Figuras 2 y 3 hemos presentado el resultado de aplicar los seis mapas de prominencia en dos imágenes de ejemplo. Visualizamos este efecto en las 40 imágenes seleccionadas y asignamos un voto positivo al mapa de prominencia únicamente cuando este contiene completamente el objeto de interés. Por ejemplo, en las Figura 2, los algoritmos Image Signature, GBVS, SDSP y SIM marcan claramente zonas prominentes dentro del objeto de interés, lo que permitiría seleccionar descriptores que pertenecerían al pájaro que está siendo descrito. Por ello, para esta imagen, dichos mapas de prominencia obtuvieron un voto positivo. Por el contrario, como puede verse en la segunda fila de dicha figura, el uso de los descriptores extraídos de las zonas marcadas por los mapas de Itti y RCSS, al no abarcar completamente el objeto de interés, daría lugar a diccionarios con palabras visuales que no incluirían algunas partes del objeto, con lo que, en este caso, no se les asignó ningún voto. En el caso de la Figura 3, GVBS y SIM conservan todos los detalles del ciempiés, con lo que se les asignó un voto positivo; en cambio, el resto de mapas no resaltan diferentes partes del objeto, por lo que no recibieron ningún voto.

Para nuestra experimentación, decidimos trabajar con los mapas GBVS y SIM, al ser los que obtuvieron las puntuaciones mayores.

3. Selección de capas y su combinación

Con los mapas de prominencia seleccionados de GBVS y SIM, nuestro objetivo consistió en evaluar en qué medida los niveles de información contenidos en diferentes capas del mapa de prominencias afectan a la clasificación de imágenes.

3.1. Creación de capas individuales

Inspirado por nuestro trabajo previo (Fidalgo et al., 2018), binarizamos el mapa de prominencia de GBVS y SIM empleando el umbral proporcionado por el método de Otsu (Otsu, 1979), obteniendo una imagen binaria donde la zona de atención corresponde con los píxeles de color blanco.

$$SM_{bin} = Otsu(SM_{\sigma}) \quad (2)$$

Considerando el mapa de prominencia como un mapa topográfico (Figura 1), realizamos varios cortes a diferentes niveles, siendo el corte superior el indicado por el valor obtenido de la Ecuación (2). De esta manera, obtenemos una imagen binaria para cada corte que utilizamos como máscara, superponiéndola a la imagen original. Esto nos permite seleccionar las características relevantes para dicho corte y enmascarar las no prominentes.

Cuando al obtener los puntos característicos de la imagen, la localización de un descriptor está dentro de una región blanca, se etiqueta como descriptor de primer plano o zona de atención; de lo contrario, se considera un descriptor de fondo.

Como se muestra en la Figura 1, cuanto más bajos realizamos los cortes en el mapa de prominencia, más información de la imagen original conservarán dichas capas. Por el contrario, cuanto más alta se encuentre la capa, mas prominente será la información recogida, aunque también puede darse el caso de que se enmascararen detalles importantes. Seleccionamos el valor de Otsu de la imagen como la capa superior para crear el mapa de prominencia binaria debido a que con el uso de un valor superior, 125 % del valor de Otsu, se creaba una máscara que apenas tenía información del objeto de interés, mientras que utilizando el umbral de Otsu se retenía gran parte de la información del objeto. Dado que los descriptores contenidos en las capas superiores también se incluyen en las inferiores, el método propuesto permite construir un diccionario para BoVWs donde los descriptores de los objetos que se encuentran en el primer plano tienen una mayor frecuencia que los descriptores de fondo o que representan objetos menos relevantes según el mapa de atención, por lo que, como se verá en los resultados obtenidos, en diversas situaciones se obtienen mejores resultados que utilizar BoVWs sobre la imagen completa.

3.2. Combinación y número de capas

Una vez que calculamos los resultados anteriores, combinamos los descriptores de los dos mapas de relevancia seleccionados, es decir, GVBS y SIM, para evaluar su contribución a la precisión final. Para ello, creamos un nuevo conjunto de descriptores resultado de la concatenación de los descriptores de primer plano extraídos de cada capa antes de crear el diccionario de palabras visuales.

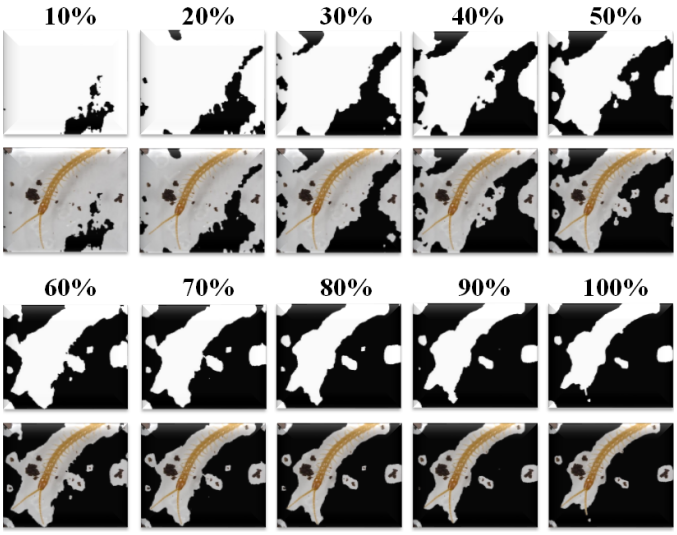

Figura 4: Esquema gráfico de la selección de características de las capas del mapa de prominencia con 10 capas. Los mapas de prominencia binarios se obtuvieron para los porcentajes del valor de Otsu indicados en la figura

Inicialmente, fusionamos los descriptores extraídos de las capas 1 y 2, dado que en la experimentación con capas individuales obtuvimos los mejores resultados para estas dos. El objetivo de esta combinación era crear un diccionario lo más discriminatorio posible. Después, concatenamos la información de las capas 1, 2, 3 y 4 para dar más peso a las características de prominencia detectadas en cada mapa. En esta fase, también probamos la concatenación de los descriptores de las capas 1 y 2, y las capas 1, 2, 3 y 4 de ambos mapas de prominencia. Estos resultados se muestran con el nombre de GBVS-SIM(1-2) y GBVS-SIM(1-4) en la Tabla 2 resumen de los resultados obtenidos. Por último, evaluamos cómo afectan las características extraídas de estos mapas de prominencia cuando el número de cortes aumenta a diez, realizando cortes en el mapa de prominencia a diez niveles equidistantes diferentes, en todos los casos utilizando el valor de Otsu como referencia.

En la Figura 4 se presenta un ejemplo de las diez capas generadas para una imagen de ImageNet-7Arthropods.

4. Experimentación

4.1. Conjuntos de datos utilizados

La Tabla 1 resume las características de los cinco conjuntos de datos utilizados, y la Figura 5 muestra algunos ejemplos, ordenados por filas, de las imágenes de cada dataset.

Tabla 1: Conjuntos de datos utilizados en la experimentación

DATASETS	Imágenes	Clases	Entrenamiento/Pruebas
Soccer	280	7	62.5/37.5
Birds	600	6	75/25
ImageNet-7Arthropods	1400	7	75/25
ImageNet-7Birds	1400	7	75/25
TOIC	698	7	75/25

Figura 5: Soccer (primera fila), Birds (segunda fila), ImageNet-7Arthropods (tercera fila), ImageNet-7Birds (cuarta fila) y TOIC (quinta fila)

Tres de los conjuntos de datos usados pueden considerarse pequeños pero difíciles: *Birds* (Lazebnik et al., 2005), con seis especies de aves en su entorno natural; *Soccer* (van de Weijer and Schmid, 2006), contiene siete categorías que corresponden a diferentes equipos de fútbol que se mezclan con otras características prominentes, como el público, la pelota o el árbitro; *TOIC* (Fidalgo et al., 2017), contiene los productos más comunes vendidos en la red oscura Tor (Al-Nabki et al., 2017) agrupados en cinco categorías.

Para verificar la validez de nuestra experimentación en conjuntos de datos más grandes, utilizamos el subconjunto *ImageNet-7Arthropods* propuesto por Fidalgo et al. (2018) y seleccionamos un nuevo conjunto de ImageNet (Russakovsky et al., 2015) al que denominamos *ImageNet-7Birds*. Ambos conjuntos tratan de reproducir las condiciones de *Birds*, pero con un conjunto de datos reciente y más exigente.

4.2. *Configuración experimental*

El método propuesto fue desarrollado en MATLAB, tomándose como referencia los resultados obtenidos con el modelo del BoVW (Csurka et al., 2004) y descriptores SIFT extraídos densamente (Lowe, 2004) y calculados sobre toda la imagen. Para los conjuntos de datos Soccer, Birds, TOIC e ImageNet-7Arthropods se han tomado como resultados de referencia los publicados en nuestros trabajos previos (Fidalgo et al., 2016, 2017; Gangwar et al., 2017; Biswas et al., 2017; Fidalgo et al., 2018).

El proceso seguido consistió en calcular inicialmente los mapas de relevancia de GBVS y SIM, para después binarizarlos y obtener las combinaciones explicadas en la sección 3. Hemos utilizado la librería VLFeat (Vedaldi and Fulkerson, 2010) para calcular los descriptores SIFT extraídos densamente de toda la imagen con paso y tamaño 7, filtrándose mediante las máscaras los mapas de prominencia analizadas en cada capa. Usamos un diccionario de 2048 palabras y construimos los vectores de características de BoVW con una codificación “dura”.

Cada conjunto de datos se dividió en dos subconjuntos, entrenamiento y pruebas, según los porcentajes indicados en la Tabla 1, utilizándose en la clasificación una Máquina de Vectores de Soporte (SVM, del inglés, *Support Vector Machine*), (Vapnik, 2000; Cervantes et al., 2017) con un kernel lineal.

5. Discusión de resultados

5.1. *Evaluación de capas individuales*

En la Figura 6 se observa cómo utilizando únicamente los descriptores obtenidos mediante el modelo de BoVW a través del filtrado realizado por algunas capas en los mapas de prominencia de GBVS y SIM, se superan los resultados de referencia. También puede observarse dos tipos de comportamiento. En el primero, que se puede apreciar en Birds e ImageNet-7Arthropods, los resultados procedentes de filtrar los descriptores con cualquiera de las cuatro capas de los mapas GBVS y SIM superan los resultados de referencia. En ambos conjuntos de datos, GBVS mejora 14.80 y 10.57 puntos, con una precisión del 82,27 ± 2,24 % y 48,86 ± 2,44 %, respectivamente.

En cambio, se observa también que tanto SIM en Soccer, GBVS en TOIC, y ambos SIM y GBVS en ImageNet-7Birds no obtuvieron una precisión superior a la de referencia en todas las capas. Puede verse que la capa 1 obtiene la mejor precisión, aumentando 3.05, 3.71 y 1.79 los resultados de referencia, con 51,05 ± 4,55 % y 73,33 ± 1,57 % en Soccer e ImageNet-7Birds con el mapa GBVS, y con 87,17 ± 4,20 % en TOIC con SIM. Por ello consideraremos que, en general, las capas de GBVS 1 y 2 son las que retienen la mayor cantidad de información del objeto de interés.

La Figura 7 representa una muestra del comportamiento de los mapas de prominencia GVBS y SIM después de extraer los diferentes niveles de información. Usaremos esta Figura para constatar los resultados obtenidos en los escenarios descritos anteriormente.

Por un lado, en la Figura 7 podemos observar cómo en ambos ejemplos, aunque todas las capas del mapa de prominencia de SIM contienen el objeto a detectar, también contienen una parte importante del fondo. Por otro lado, solo las capas uno y dos de GBVS contienen de forma completa el objeto de interés y tienen menos información de fondo que SIM. Esa es la razón por la que las secciones de GBVS uno y dos obtienen las mayores mejoras sobre los resultados de referencia. Para terminar, se incorporan los resultados de la Figura 6 en la Tabla 2, lo que permitirá al lector comparar los resultados de la Secciones 5.1 y 5.2.

5.2. *Evaluación de las combinaciones de las capas y de su número*

En esta sección, presentamos los resultados de las combinaciones descritas en la Sección 3.2. Hemos denominado GBVS(1-2) y SIM(1-2) a los resultados obtenidos utilizando el conjunto de descriptores provenientes de las dos primeras capas de cada mapa de prominencia, GBVS(1-4) y SIM(1-4) a la combinación de la información de primer plano de las cuatro capas analizadas, GBVS-SIM(1-2) y GBVS-SIM(1-4) a la concatenación de las dos primeras y las cuatro capas de ambos mapas de prominencia y GBVS (1-10) y SIM (1-10) a la combinación de las diez capas extraídas de un mismo mapa.

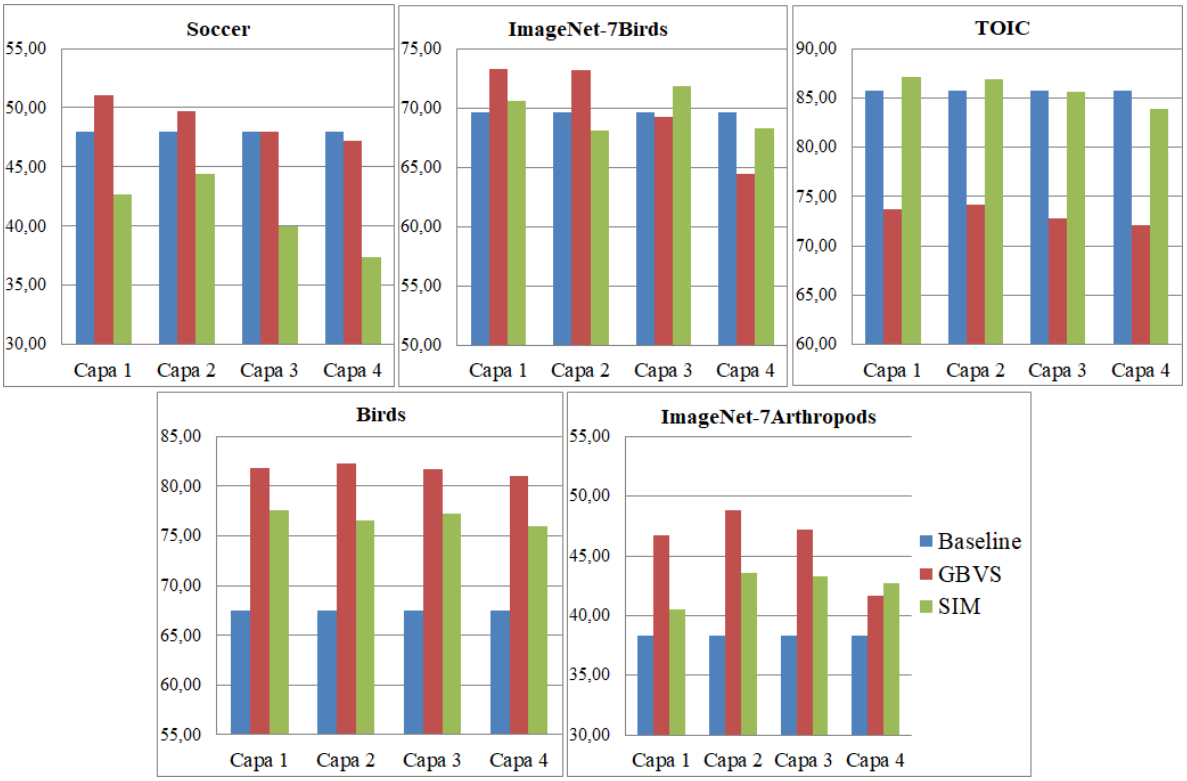

Figura 6: Precisión obtenida en la evaluación por capas individuales para cada conjunto de datos y su comparación con los resultados de referencia o Baseline

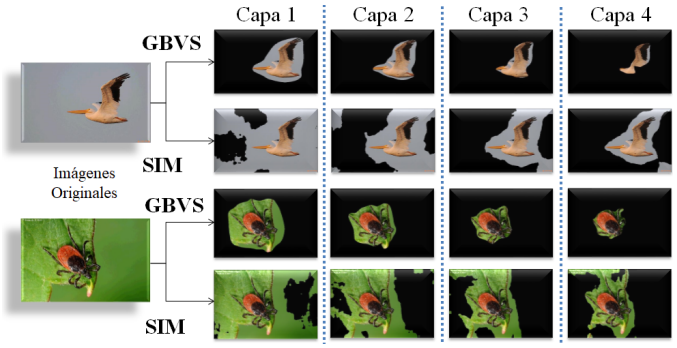

Figura 7: Ejemplo de capas extraídas de los mapas de prominencia GBVS y SIM en los conjuntos de datos ImageNet-7Birds e ImageNet-7Arthropods

En la Figura 8 se puede comprobar como todos los resultados obtenidos superan a los resultados de referencia, excepto las combinaciones de GBVS en el conjunto de datos TOIC y Soccer. Debido a la complejidad de ambos conjuntos de datos, donde las imágenes presentan mucha información prominente no relacionada con la clase de la imagen, es decir, con la categoría ilegal en TOIC y la identificación del equipo de fútbol, respectivamente. Ello provoca que las regiones binarias de GBVS de las diferentes capas retengan información de elementos no relacionados con dicha clase, penalizando los resultados obtenidos. Excepto en el conjunto de datos de Soccer y TOIC con GBVS, podemos concluir que las diferentes combinaciones de capas implican una mejora en la precisión en comparación con los resultados de referencia.

Adicionalmente, queremos resaltar nuestras observaciones sobre el número de capas y la combinación de dichas capas entre los dos mapas de prominencia. Por un lado, cuando se combinan capas en el mapa de prominencia de SIM se mejora la precisión obtenida al aumentar el número de cortes de dos a diez, es decir, SIM(1-2), SIM(1-4) y SIM(1-10), en los conjuntos de datos Soccer, Birds y TOIC. Sin embargo, en los subconjuntos de ImageNet mejora en SIM(1-2) y SIM(1-4), pero disminuye la precisión significativamente al aumentar el número de capas a 10. Queremos resaltar que la combinación de las dos primeras capas, es decir, SIM(1-2) y GBVS(1-2) obtiene una mayor precisión en todos los casos si se comparan con el uso de la información extraída de cada capa por separado, excepto en el conjunto de datos Soccer. La explicación puede encontrarse nuevamente en el ejemplo mostrado en la Figura 7. Como estamos seleccionando las capas que retendrían todo el primer plano y descartando principalmente la información de fondo, el diccionario agrupado contendría palabras visuales óptimas para que la codificación posterior sea más eficiente.

Por otro lado, la combinación de la información extraída de las capas de GBVS no presenta un comportamiento uniforme cuando aumenta el número de capas, pero obtiene la mejor precisión para los conjuntos de datos ImageNet-7Arthropods, ImageNet-7Birds, con precisiones de 50.66±5.62 % y 76.00±2.00 %, lo que representa 12.37 y 6.48 puntos de mejora con respecto a sus resultados de referencia. Cuando se combina información de capas procedentes de los dos mapas de relevancia, es decir, GBVS+SIM(1-2) y GBVS+SIM(1-4), se obtienen los mejores resultados para los conjuntos de datos de Soccer, Birds y TOIC, con 52.38±3.93 %, 84.27±1.01 % y 88.32±4.26 %, es decir, una precisión de 4.38, 16.80 y 2.54 puntos superior, respectivamente. En estos tres conjuntos de datos, el hecho de que las dos primeras capas re-

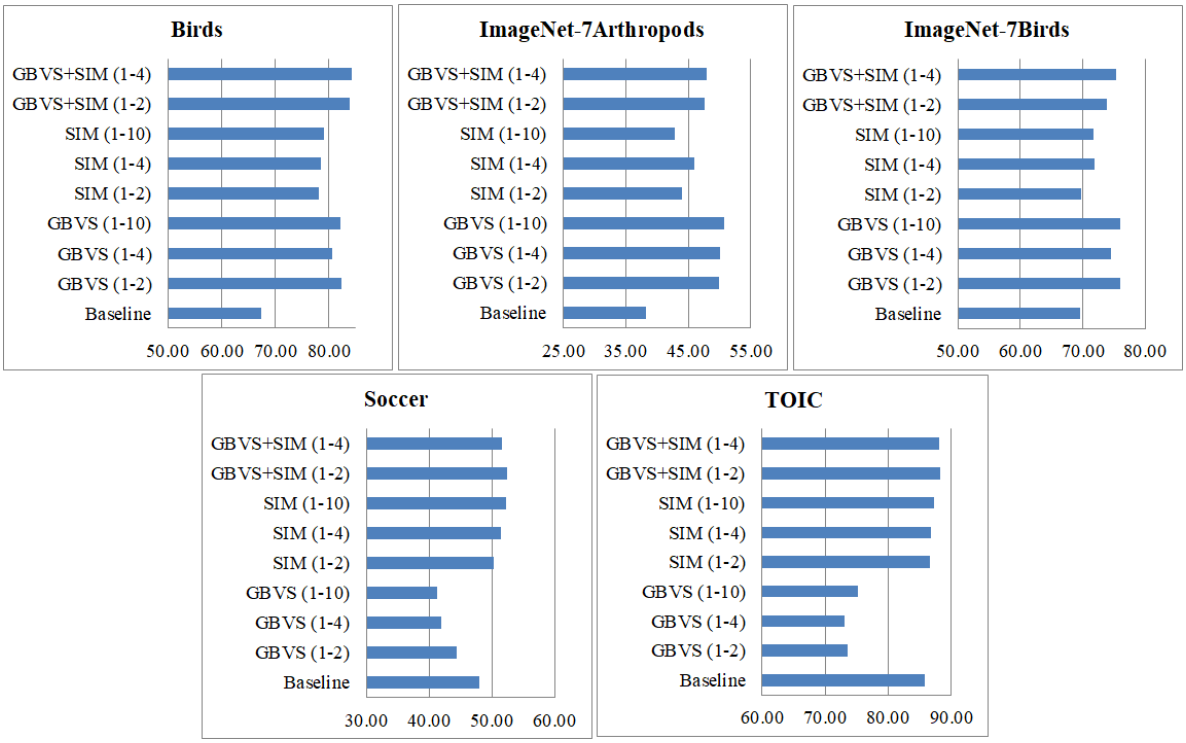

Figura 8: Resultados de las diferentes combinaciones de capas: (1-2) primeras dos capas; (1-4) primeras cuatro capas; (1-10) diez capas; X+X(1-2) combinación de las dos primeras capas de ambos mapas de prominencia; X+X(1-4) combinación de las cuatro capas.

tengan información de los objetos de interés da como resultado una mejor precisión. Al incluir la información de la tercera y cuarta capa de GBVS y SIM se obtiene el mejor resultado para Birds, pero no para Soccer y TOIC con solo las dos primeras capas.

Para finalizar, en la Tabla 2 se presentan de forma resumida todos los resultados obtenidos en la experimentación realizada, lo que permite revisar los resultados comentados.

6. Conclusiones y trabajos futuros

El objetivo de los mapas de prominencia es resaltar las áreas de atención de una imagen ayudando de esta forma a localizar los objetos de interés. La información contenida en un mapa de prominencia podría representarse como una superficie topográfica con diferentes niveles de información. En este trabajo, inicialmente calculamos los descriptores SIFT extraídos densamente para toda la imagen. Hemos demostrado que utilizando cuatro niveles diferentes de este área topográfica, que hemos denominado capas, es posible seleccionar un conjunto de descriptores que mejoran la precisión en la clasificación de imágenes. Después de binarizar las cuatro capas extraídas del mapa de prominencia, utilizamos la máscara resultante para seleccionar un subconjunto de los descriptores SIFT y crear un modelo de BoVW. Al analizar la precisión obtenida por los descriptores de cada capa en la clasificación de imágenes, utilizando para ello máquinas de vectores de soporte, hemos observado que si la capa retiene la mayoría de la información relacionada con el objeto, la precisión obtenida es mayor que con el método de referencia, es decir, usando todos los descriptores SIFT de toda la imagen. Cuando esto no sucede, los descriptores extraídos no retienen información suficiente para superar los resultados de referencia.

Después de evaluar la información retenida por cada capa en base a cómo afecta a la clasificación, decidimos fusionar los descriptores de forma temprana. Esta fusión se basa en adicionar al modelo los descriptores filtrados en cada una de las capas consideradas, lo que produce un efecto de ponderación de las características -sus descriptores- presentes en más de una capa, antes de la generación del diccionario visual en el modelo BoVW. De esta manera, obtuvimos una codificación optimizada del vector de características en el BoVW y conseguimos superar la precisión obtenida individualmente por los descriptores de cada capa.

Dentro de un mismo mapa de prominencia, combinamos los descriptores de las dos y de las cuatro primeras capas, respectivamente. También evaluamos los resultados obtenidos al dividir el mapa de prominencia en diez capas combinándolas posteriormente. Excepto en los conjuntos de datos de Soccer y TOIC, podemos concluir que las diferentes combinaciones de capas analizadas obtienen una precisión superior al método base, donde se crea el diccionario con descriptores provenientes de la imagen completa. Nuestra propuesta funciona peor en estos dos datasets porque ambos conjuntos presentan múltiples objetos prominentes dentro de una misma imagen, algunos de los cuales no pertenecen a la categoría a clasificar. Dicha situación implica que las diferentes capas almacenan información no relevante que penaliza los resultados. En los dos subconjuntos de datos procedentes de ImageNet, obtuvimos la mejor precisión cuando combinamos las diez capas extraídas del mapa de prominencia GBVS. En base a estos experimentos, hemos determinado que al aumentar el número de capas que se combinan, desde dos hasta diez, para el mapa de prominencia de

Tabla 2: Resumen de los resultados obtenidos en los cinco conjuntos de datos. *Baseline* se refiere a los resultados de referencia

Experimento	Soccer	Birds	ImageNet-7Arthropods	ImageNet-7Birds	TOIC
Baseline	48.00±3.66 %	67.47±2.96 %	38.29±1.63 %	69.62±1.44 %	85.78±4.25 %
SIM Capa 1	42.67±2.81 %	77.60±1.74 %	40.48±2.88 %	70.67±3.73 %	87.17±4.20 %
SIM Capa 2	44.38±4.40 %	76.53±2.08 %	43.52±1.15 %	68.10±1.57 %	86.94±5.76 %
SIM Capa 3	40.00±3.63 %	77.20±1.79 %	43.24±1.29 %	71.90±1.00 %	85.66±4.79 %
SIM Capa 4	37.33±4.73 %	76.00±3.56 %	42.67±2.59 %	68.29±1.48 %	83.93±4.52 %
SIM (1-2)	50.29±4.78 %	78.13±2.47 %	44.00±1.25 %	69.81±3.43 %	86.59±5.16 %
SIM (1-4)	51.43±4.57 %	78.53±3.75 %	45.90±2.59 %	71.90±2.07 %	86.82±4.16 %
SIM (1-10)	52.19±5.06 %	79.20±3.84 %	42.76±1.67 %	71.71±2.62 %	87.28±4.70 %
GBVS Capa 1	51.05±4.55 %	81.87±0.73 %	46.76±2.01 %	73.33±1.57 %	73.76±5.43 %
GBVS Capa 2	49.71±3.77 %	82.27±2.24 %	48.86±2.44 %	73.24±1.57 %	74.22±3.71 %
GBVS Capa 3	48.00±1.28 %	81.73±1.01 %	46.19±2.43 %	69.24±1.47 %	73.18±4.22 %
GBVS Capa 4	47.24±6.05 %	81.07±3.90 %	47.24±0.72 %	64.48±1.72 %	74.22±4.24 %
GBVS (1-2)	44.38±4.02 %	82.40±1.53 %	50.00±2.00 %	76.00±2.00 %	73.64±4.36 %
GBVS (1-4)	41.90±0.67 %	80.67±2.45 %	50.10±0.87 %	74.57±1.74 %	73.18±4.22 %
GBVS (1-10)	41.33±4.79 %	82.13±3.18 %	**50.66±5.62 %**	**76.00±1.03 %**	75.26±3.82 %
GBVS+SIM (1-2)	**52.38±3.93 %**	83.87±2.18 %	47.65±2.95 %	73.90±2.78 %	**88.32±4.26 %**
GBVS+SIM (1-4)	51.62±1.95 %	**84.27±1.01 %**	47.90±1.08 %	75.34±0.83 %	88.09±3.13 %

SIM se produce un aumento en la precisión. En cambio, para el mapa de prominencia de GBVS, la combinación de capas no presenta un comportamiento uniforme en cuanto a la precisión resultante. Por último, cuando hemos combinado las dos y cuatro primeras capas de los mapas de atención de GBVS y SIM, hemos obtenido los mejores resultados en los datasets de Soccer, Birds y TOIC.

En resumen, podemos concluir que la fusión temprana de descriptores SIFT, utilizando capas obtenidas según el modelo propuesto de los mapas de prominencia GBVS y SIM mejoran la clasificación de imágenes en conjuntos que no presentan múltiples objetos por imagen, frente a los resultados obtenidos con el modelo BoVW clásico.

En el futuro, trabajaremos en la selección automática de las capas que pueden contener la mayor parte de la información del objeto de interés y en la aplicación de esta propuesta en conjuntos de datos con múltiples objetos por imagen.

Agradecimientos

Este trabajo ha sido financiado por el contrato INCIBE "INCIBEI-2015-27359" correspondiente a las "Ayudas para la Excelencia de los Equipos de Investigación avanzada en ciberseguridad" y por el acuerdo del convenio marco firmado entre la Universidad de León y el INCIBE (Instituto Nacional de Ciberseguridad) bajo la Adenda 22.

Referencias

Al-khafaji, S. L., Zhou, J., Zia, A., Liew, A. W. C., Feb 2018. Spectral-spatial scale invariant feature transform for hyperspectral images. IEEE Transactions on Image Processing 27 (2), 837–850.
DOI: 10.1109/TIP.2017.2749145

Al-Nabki, W., Fidalgo, E., Alegre, E., De Paz, I., 2017. Classifying Illegal Activities on Tor Network Based on Web Textual Contents. 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference 1, 35–43.

Beucher, S., Lantuejoul, C., 1979. Use of Watersheds in Contour Detection.
DOI: citeulike-article-id:4083187

Biagio, M. S., Bazzani, L., Cristani, M., Murino, V., oct 2014. Weighted bag of visual words for object recognition. In: 2014 IEEE International Conference on Image Processing, ICIP 2014. IEEE, pp. 2734–2738.
DOI: 10.1109/ICIP.2014.7025553

Biswas, R., Fidalgo, E., Alegre, E., 2017. Recognition of Service Domains on TOR Dark Net using Perceptual Hashing and Image Classification Techniques. 8th International Conference on Imaging for Crime Detection and Prevention 2017 (5).
DOI: 10.1049/ic.2017.0041

Borji, A., Itti, L., jan 2013. State-of-the-art in visual attention modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (1), 185–207.
DOI: 10.1109/TPAMI.2012.89

Cervantes, J., Taltempa, J., García-Lamont, F., Castilla, J. S. R., Rendon, A. Y., Jalili, L. D., 2017. Anélisis comparativo de las técnicas utilizadas en un sistema de reconocimiento de hojas de planta. Revista Iberoamericana de Automética e Informética Industrial RIAI 14 (1), 104 –114.

Chatzichristofis, S. A., Iakovidou, C., Boutalis, Y., Marques, O., feb 2013. Co.Vi.Wo.: Color visual words based on non-predefined size codebooks. IEEE Transactions on Cybernetics 43 (1), 192–205.
DOI: 10.1109/TSMCB.2012.2203300

Chaves, D., Saikia, S., Fernández-Robles, L., Alegre, E., Trujillo, M., 2018. A Systematic Review on Object Localisation Methods in Images. Revista Iberoamericana de Automática e Informática Industrial 15, 231–242.
DOI: https://doi.org/10.4995/riai.2018.10229

Chen, J., Feng, B., Xu, B., 2014a. Spatial similarity measure of visual phrases for image retrieval. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Vol. 8326 LNCS. Springer, Cham, pp. 275–282.
DOI: 10.1007/978-3-319-04117-9_25

Chen, Y., Li, X., Dick, A., Hill, R., 2014b. Ranking consistency for image matching and object retrieval. Pattern Recognition 47 (3), 1349 – 1360, handwriting Recognition and other PR Applications.
DOI: https://doi.org/10.1016/j.patcog.2013.09.011

Csurka, G., Csurka, G., Dance, C. R., Fan, L., Willamowski, J., Bray, C., 2004. Visual categorization with bags of keypoints. IN WORKSHOP ON STATISTICAL LEARNING IN COMPUTER VISION, ECCV, 1—-22.

Digabel, H., Lantuéjoul, C., 1978. Iterative algorithms. Actes du Second Symposium Européen d'Analyse Quantitative des Microstructures en Sciences des Matériaux, Biologie et Médecine 1 (1), 85–99.

Fang, Y., Lei, J., Li, J., Xu, L., Lin, W., Callet, P. L., 2017. Learning visual saliency from human fixations for stereoscopic images. Neurocomputing 266, 284 – 292.
DOI: https://doi.org/10.1016/j.neucom.2017.05.050

Fidalgo, E., Alegre, E., González-Castro, V., Fernández-Robles, L., 2016. Compass radius estimation for improved image classification using Edge-SIFT. Neurocomputing 197, 119–135.
DOI: 10.1016/j.neucom.2016.02.045

Fidalgo, E., Alegre, E., González-Castro, V., Fernández-Robles, L., 2017. Illegal activity categorisation in DarkNet based on image classification using CREIC method. 10th International Conference on Computational Intelligence in Security for Information Systems I (1), 600–609.

DOI: 10.1007/978-3-319-67180-2_58
Fidalgo, E., Alegre, E., González-Castro, V., Fernández-Robles, L., 2018. Boosting image classification through semantic attention filtering strategies. Pattern Recognition Letters.
DOI: 10.1016/j.patrec.2018.06.033
Field, D. J., dec 1987. Relations between the statistics of natural images and the response properties of cortical cells. Journal of the Optical Society of America. A, Optics and image science 4 (12), 2379–94.
DOI: 10.1364/JOSAA.4.002379
Gangwar, A., Fidalgo, E., Alegre, E., González-Castro, V., Dec 2017. Pornography and child sexual abuse detection in image and video: A comparative evaluation. In: 8th International Conference on Imaging for Crime Detection and Prevention (ICDP 2017). pp. 37–42.
DOI: 10.1049/ic.2017.0046
García-Olalla, O., Alegre, E., Fernández-Robles, L., Fidalgo, E., Saikia, S., 2018. Textile retrieval based on image content from cdc and webcam cameras in indoor environments. Sensors 18 (5).
DOI: 10.3390/s18051329
Gonzalez, R., Woods, R., 2002. Digital image processing. Prentice Hall.
DOI: 10.1016/0734-189X(90)90171-Q
González-Castro, V., Valdés Hernández, M. d. C., Chappell, F. M., Armitage, P. A., Makin, S., Wardlaw, J. M., 2017. Reliability of an automatic classifier for brain enlarged perivascular spaces burden and comparison with human performance. Clinical Science 131 (13), 1465–1481.
DOI: 10.1042/CS20170051
Greenspan, H., Belongie, S., Goodman, R., Perona, P., Rakshit, S., Anderson, C., 1994. Overcomplete steerable pyramid filters and rotation invariance. Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on, 0–6.
DOI: 10.1109/CVPR.1994.323833
Harel, J., Koch, C., Perona, P., 2007. Graph-Based Visual Saliency.
He, Y., Deng, G., Wang, Y., Wei, L., Yang, J., Li, X., Zhang, Y., 2018. Optimization of sift algorithm for fast-image feature extraction in line-scanning ophthalmoscope. Optik 152, 21 – 28.
DOI: https://doi.org/10.1016/j.ijleo.2017.09.075
Hou, X., Harel, J., Koch, C., 2012. Image signature: Highlighting sparse salient regions. IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (1), 194–201.
DOI: 10.1109/TPAMI.2011.146
Itti, L., Koch, C., Niebur, E., 1998. A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (11), 1254–1259.
DOI: 10.1109/34.730558
Jian, M., Wu, L., Jung, C., Fu, Q., Jia, T., 2018. Visual saliency estimation using constraints. Neurocomputing 290, 1 – 11.
DOI: 10.1016/j.neucom.2018.02.004
Kienzle, W., Wichmann, F., Schölkopf, B., Franz, M., 2007. A Nonparametric Approach to Bottom-Up Visual Saliency. Advances in Neural Information Processing Systems 19 19 (December 2006), 689–696.
Lahouli, I., Karakasis, E., Haelterman, R., Chtourou, Z., Cubber, G. D., Gasteratos, A., Attia, R., 2018. Hot spot method for pedestrian detection using saliency maps, discrete chebyshev moments and support vector machine. IET Image Processing 12 (7), 1284–1291.
DOI: 10.1049/iet-ipr.2017.0221
Lazebnik, S., Schmid, C., Ponce, J., 2005. A maximum entropy framework for part-based texture and object recognition. Proceedings of the IEEE International Conference on Computer Vision I (1), 832–838.
DOI: 10.1109/ICCV.2005.10
Lowe, D. G., 2004. Distinctive image features from scale invariant keypoints. Int'l Journal of Computer Vision 60, 91–11020042.
DOI: http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94
Mallat, S., mar 2009. Geometrical grouplets. Applied and Computational Harmonic Analysis 26 (2), 161–180.
DOI: 10.1016/j.acha.2008.03.004
Margolin, R., Zelnik-Manor, L., Tal, A., may 2013. Saliency for image manipulation. Visual Computer 29 (5), 381–392.
DOI: 10.1007/s00371-012-0740-x
Murray, N., Vanrell, M., Otazu, X., Parraga, C. A., nov 2013. Low-level spatiochromatic grouping for saliency estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (11), 2810–2816.
DOI: 10.1109/TPAMI.2013.108
Otsu, N., 1979. A threshold selection method from Gray-level. IEEE Transactions on Systems, Man, and Cybernetics SMC-9 (1), 62–66.
DOI: 10.1109/TSMC.1979.4310076
Pinto, N., Doukhan, D., DiCarlo, J. J., Cox, D. D., nov 2009. A high-throughput screening approach to discovering good forms of biologically inspired visual representation. PLoS Computational Biology 5 (11), e1000579.
DOI: 10.1371/journal.pcbi.1000579
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., Fei-Fei, L., dec 2015. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision 115 (3), 211–252.
DOI: 10.1007/s11263-015-0816-y
Saikia, S., Fidalgo, E., Alegre, E., Fernández-Robles, L., sep 2017. Object Detection for Crime Scene Evidence Analysis Using Deep Learning. In: International Conference on Image Analysis and Processing. Springer, Cham, pp. 14–24.
DOI: 10.1007/978-3-319-68548-9_2
Saikia, S., Fidalgo, E., Alegre, E., Fernández-Robles, L., 2018. Query based object retrieval using neural codes. In: Advances in Intelligent Systems and Computing. Vol. 649. Springer, Cham, pp. 513–523.
DOI: 10.1007/978-3-319-67180-2_50
Sepúlveda, G. V., Torriti, M. T., Calero, M. F., 2017. Sistema de detección de señales de tráfico para la localización de intersecciones viales y frenado anticipado. Revista Iberoamericana de Automática e Informática Industrial 14 (2), 152–162.
DOI: 10.1016/j.riai.2016.09.010
Shen, X., Wu, Y., jun 2012. A unified approach to salient object detection via low rank matrix recovery. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, pp. 853–860.
DOI: 10.1109/CVPR.2012.6247758
Tilke, J., Ehinger, K., Durand, F., Torralba, A., sep 2009. Learning to predict where humans look. In: Proceedings of the IEEE International Conference on Computer Vision. IEEE, pp. 2106–2113.
DOI: 10.1109/ICCV.2009.5459462
Toet, A., Sadaka, N. G., Karam, jun 2009. Frequency-tuned salient region detection. Vision Research 45 (1), II – 169–II – 172.
DOI: 10.1109/CVPR.2009.5206596
Trzcinski, T., Christoudias, M., Lepetit, V., mar 2015. Learning image descriptors with boosting. IEEE Transactions on Pattern Analysis and Machine Intelligence 37 (3), 597–610.
DOI: 10.1109/TPAMI.2014.2343961
van de Weijer, J., Schmid, C., 2006. Coloring Local Feature Extraction. In: Computer Vision – ECCV 2006. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 334–348.
DOI: 10.1007/11744047_26
Vapnik, V. N., 2000. The Nature of Statistical Learning Theory. Springer New York.
DOI: 10.1007/978-1-4757-3264-1
Vedaldi, A., Fulkerson, B., 2010. Vlfeat. Proceedings of the international conference on Multimedia - MM '10 3 (1), 1469.
DOI: 10.1145/1873951.1874249
Vikram, T. N., Tscherepanow, M., Wrede, B., 2012. A saliency map based on sampling an image into random rectangular regions of interest. Pattern Recognition 45 (9), 3114 – 3124, best Papers of Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA'2011).
DOI: https://doi.org/10.1016/j.patcog.2012.02.009
Yan, Q., Xu, L., Shi, J., Jia, J., June 2013. Hierarchical saliency detection. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1155–1162.
DOI: 10.1109/CVPR.2013.153
Zhang, L., Gu, Z., Li, H., sep 2013. SDSP: A novel saliency detection method by combining simple priors. In: 2013 IEEE International Conference on Image Processing, ICIP 2013 - Proceedings. pp. 171–175.
DOI: 10.1109/ICIP.2013.6738036
Zhao, Q., Koch, C., jun 2012. Learning visual saliency by combining feature maps in a nonlinear manner using AdaBoost. Journal of Vision 12 (6), 22.
DOI: 10.1167/12.6.22
Zheng, L., Wang, S., Liu, Z., Tian, Q., jun 2013. Lp-Norm IDF for Large Scale Image Search. Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, 1626–1633.
DOI: 10.1109/CVPR.2013.213
Zheng, L., Wang, S., Zhou, W., Tian, Q., jun 2014. Bayes merging of multiple vocabularies for scalable image retrieval. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1963–1970.
DOI: 10.1109/CVPR.2014.252