

Tipo de artículo: Artículo original

Publicado: dd/mm/aa En: XIII Congreso Nacional de Reconocimiento de Patrones

Propuestas para el mejoramiento del enfoque BoW en la clasificación de objetos

On improving the BoW approach in Object Classification

Máximo Rodríguez-Collada¹, Airl Pérez-Suárez^{1*}, Leonardo Chang¹

¹Centro de Aplicaciones de Tecnologías de Avanzada (CENATAV). 7a No. 21406 e/ 214 and 216, Siboney, Playa, CP 12200, La Habana, Cuba.

*Autor para correspondencia: asuares@cenatav.co.cu

Resumen

El enfoque Bolsa de Palabras (BoW, por sus siglas en inglés) es uno de los enfoques más usados para representar imágenes en el contexto de la categorización de objetos; Sin embargo, este enfoque tiene varias limitaciones. En este trabajo, se proponen tres propiedades y sus correspondientes medidas, para evaluar cuantitativamente la habilidad que tiene una palabra visual para representar y distinguir a una clase de objeto. Adicionalmente, se introducen dos métodos para el ranking y filtrado de los vocabularios visuales, así como un nuevo método para la representación de las imágenes, a partir de estos vocabularios. Los experimentos realizados sobre el conjunto de datos Caltech-101 mostraron las mejoras introducidas por cada una de las propuestas, las cuales obtienen los mejores resultados de clasificación para las tasas de compresión más altas, en comparación con los resultados alcanzados por un método basado en Información Mutua, reportado en el estado-del-arte.

Palabras claves: Bolsa de Palabras Visuales, Vocabulario Visual, Categorización de Objetos, Reconocimiento de Objetos.

Abstract

Bag of Words (BoW) is one of the most widely used approaches for representing images for object categorization; however, it has several drawbacks. In this paper, we propose three properties and their corresponding quantitative evaluation measures to assess the ability of a visual word to represent and discriminate an object class. Additionally, we also introduce two methods for ranking and filtering visual vocabularies and a soft weighting method for BoW image representation. Experiments conducted on the Caltech-101 dataset showed the improvement introduced by our proposals, which obtained the best classification results for the highest compression rates when compared with a state-of-the-art mutual information based method for feature selection.

Keywords: *Bag of Visual Words, Visual Vocabulary, Object Categorization, Object Recognition.*

Introducción

Uno de los enfoques más usados para la representación de imágenes en el contexto de la categorización de objetos, es el de Bolsa de Palabras (BoW, por sus siglas en inglés). Los métodos basados en este enfoque han obtenido notables resultados en años recientes y han alcanzado los mejores resultados para varias clases en la más reciente competencia PASCAL VOC de clasificación de objetos.

La idea central del enfoque BoW es discretizar el espacio de descriptores locales (e.g., SIFT ([Lowe, 2004](#))) extraídos de un conjunto de imágenes de entrenamiento, en puntos de interés o muestrados densamente en las imágenes. Para esto, se ejecuta un algoritmo de agrupamiento sobre el conjunto de descriptores extraídos de todas las imágenes de entrenamiento, con el objetivo de identificar grupos de descriptores que son visualmente equivalentes. Cada grupo se interpreta como una *palabra visual* y el conjunto de todos los grupos forma el llamado *vocabulario visual*. Posteriormente, para representar a una imagen del conjunto de entrenamiento, a cada uno de sus descriptores se le asigna una palabra visual y a partir de estas asignaciones se construye un *histograma de ocurrencias* de las palabras visuales en la imagen; este histograma constituye la representación de dicha imagen. Luego, cuando se desea clasificar una imagen no vista, esta es representada utilizando el vocabulario visual y es procesada por el clasificador.

Una de las principales limitaciones del enfoque BoW es que en la construcción del vocabulario visual se utilizan descriptores del objeto contenido en la imagen, así como del fondo de las mismas. Esto implica que el ruido extraído del fondo de las imágenes es considerado también como parte de la descripción de la clase del objeto. Adicionalmente, en el proceso de representación de las imágenes, cada palabra visual es utilizada y contribuye en igual magnitud a la construcción del histograma, independientemente de qué tan bien esta palabra represente y distinga a esta imagen del resto. Todos los elementos comentados en este párrafo pueden reducir la eficacia del proceso de clasificación.

En la literatura se han propuesto varios métodos para atacar las limitaciones del enfoque BoW. Estas propuestas incluyen trabajos recientes que se enfocan en construir vocabularios visuales más representativos y distintivos, como por ejemplo los trabajos ([Kesorn and Poslad, 2012](#); [Lopez-Sastre et al., 2011](#); [Jiu et al., 2012](#)), en los cuales los autores usan la etiqueta de las clases de las imágenes, en la etapa de entrenamiento, para obtener mejores vocabularios. Por otro lado, dado que cuando se construye el histograma de ocurrencias de las imágenes utilizando el procedimiento estándar, el vocabulario resultante no permite representar bien aquellos descriptores que se encuentran cerca de los límites de Voronoi, los investigadores se han enfocado recientemente en proponer métodos que realizan múltiples asignaciones o asignaciones difusas. Por ejemplo, los trabajos ([Chang et al., 2012](#); [Jiang et al., 2007](#)) proponen métodos que realizan asignaciones múltiples, en los cuales un descriptor es asociado con sus k palabras visuales más cercanas. El trabajo que más se relaciona

con nuestra propuesta es el reciente trabajo de Zhang *et al.* (Zhang et al., 2014), en el que se propone un método supervisado basado en Información Mutua (MI, por sus siglas en inglés). Este método utiliza MI entre cada dimensión del descriptor de la imagen y la etiqueta de clase de la imagen, con el objetivo de calcular la importancia de cada dimension. Finalmente, utilizando las dimensiones más importantes logran reducir el tamaño de la representación de la imagen. Este método alcanza eficacias más altas que métodos basados en compresión de descriptores tales como PQ (Jégou et al., 2011) y BPBC (Gong et al., 2013).

En este trabajo, se presentan tres propiedades que permiten evaluar la idoneidad de una palabra visual para representar y distinguir a una clase de objeto, en el contexto del enfoque BoW. Se definen tres medidas con el objetivo de evaluar cuantitativamente cada una de estas tres propiedades y se proponen dos métodos de ranking y filtrado del vocabulario visual, uno basado en las medidas comentadas y otro basado en un esquema de pesado *tf.idf*. Por último, se propone también una alternativa para la representación de las imágenes a partir del vocabulario visual, que tiene en cuenta el poder descriptivo y distintivo de cada palabra visual. Los experimentos realizados sobre el conjunto Caltech-101 (Fei-Fei et al., 2006), en la tarea de clasificación, muestran las mejoras introducidas por nuestras propuestas, las cuales alcanzan eficacias más altas para los mayores índices de compresión, en relación con el método de Zhang *et al.* (Zhang et al., 2014).

Métodos propuestos

En esta sección se proponen tres propiedades que debiera cumplir una palabra visual para ser representativa y distintiva de una clase de objetos. También se presentan las medidas que permiten evaluar cuantitativamente el grado de cumplimiento de estas propiedades por parte de una palabra visual. Además, se proponen dos métodos de ranking y filtrado del vocabulario visual y un método para la representación de las imágenes a partir del vocabulario visual, que está basado en las medidas anteriormente comentadas.

Medida de Representatividad Inter-clase

Una palabra visual puede estar compuesta por descriptores de diferentes clases de objetos, representando conceptos visuales que son comunes en estas clases diferentes. A su vez, estas partes en común no tienen que estar necesariamente igual representadas dentro de la palabra visual, ya que incluso siendo semejantes, las clases de objetos a las que pertenecen también tienen partes que las distinguen. Con base en esto, se puede decir que una palabra visual representa mejor a una clase de objeto mientras mayor es la representatividad de esta clase dentro de la palabra. Para medir la representatividad de una clase c_j en la palabra visual k , se propone la medida \mathcal{M}_1 :

$$\mathcal{M}_1(k, c_j) = \frac{f_{k, c_j}}{n_k}, \quad (1)$$

donde f_{k,c_j} representa el número de descriptores de la clase c_j presentes en la palabra visual k y n_k es el número total de descriptores presentes en la palabra visual k .

Medida de Representatividad Intra-clase

Una palabra visual puede estar compuesta por descriptores de diferentes objetos, muchos de ellos posiblemente pertenecientes a la misma clase de objetos. Incluso siendo diferentes, los objetos de una misma clase deben compartir varios conceptos visuales. Tomando esto en cuenta, se puede decir que una palabra visual describirá mejor a una clase de objetos mientras más balanceada sea la presencia de los descriptores de dicha clase dentro de la palabra visual, en relación al número total de objetos que existan en dicha clase en el conjunto de entrenamiento. Por lo tanto, se puede decir que una palabra visual representa bien a una clase si tiene una alta representatividad intra-clase para dicha clase. Para medir la representatividad intra-clase que tiene la palabra visual k para la clase c_j , se propone la medida μ :

$$\mu(k, c_j) = \frac{1}{O_{c_j}} \sum_{m=1}^{O_{c_j}} \left| \frac{o_{m,k,c_j}}{f_{k,c_j}} - \frac{1}{O_{c_j}} \right|, \quad (2)$$

donde O_{c_j} es el número de objetos (imágenes) de la clase c_j en el conjunto de entrenamiento. o_{m,k,c_j} es el número de descriptores del objeto m , perteneciente a la clase c_j , que están incluidos en la palabra visual k , y f_{k,c_j} es el número total de descriptores de la clase c_j presentes en la palabra visual k . La razón que garantiza el mejor balance de la clase c_j está determinada por el término $1/O_{c_j}$, que representa el caso en que cada objeto de la clase c_j está igualmente representado dentro de la palabra visual k .

La medida μ evalúa cuánto una clase de objetos dada se desvía de su valor ideal de balance intra-clase. Con el objetivo de hacer este valor comparable con otras clases y palabras visuales, μ debe ser normalizada usando su valor máximo posible, el cual es $\frac{2 \cdot O_{c_j} - 2}{O_{c_j}^2}$. Tomando en cuenta que μ alcanza este máximo valor en el caso peor de representatividad intra-clase, se define la medida \mathcal{M}_2 de forma tal que esta esté normalizada por $\max(\mu(k, c_j))$ y alcance su máximo valor posible cuando se alcance el valor ideal de balance intra-clase variability:

$$\mathcal{M}_2(k, c_j) = 1 - \frac{O_{c_j}}{2 \cdot (O_{c_j} - 1)} \sum_{m=1}^{O_{c_j}} \left| \frac{o_{m,k,c_j}}{s_{k,c_j}} - \frac{1}{O_{c_j}} \right|. \quad (3)$$

Medida de distintividad inter-clase

\mathcal{M}_1 y \mathcal{M}_2 brindan, bajo perspectivas distintas, una evaluación cuantitativa acerca de la habilidad de una palabra visual para describir una clase de objetos dada. Sin embargo, un vocabulario visual no debe construirse

seleccionando solamente aquellas palabras visuales que mejor representen a cada clase de objeto, ya que este hecho no garantiza que dichas palabras sean capaces de distinguir bien una clase de otra, como se espera haya un vocabulario visual. Por lo tanto, se puede afirmar que, con el objetivo de ser utilizada como parte de un vocabulario visual, una propiedad deseada en una palabra visual es que tenga altos valores de $\mathcal{M}_1(k, c_j)$ y $\mathcal{M}_2(k, c_j)$ (represente bien una clase de objetos) y bajos valores de $\mathcal{M}_1(k, \{c_j\}^C)$ and $\mathcal{M}_2(k, \{c_j\}^C)$ (no represente bien el resto de las clases); es decir, la palabra debe tener alto poder distintivo.

Para poder ser capaces de cuantificar el poder distintivo de una palabra visual para una clase de objetos dada, se propone la medida \mathcal{M}_3 . Esta medida expresa cuánto se separa del resto de las clases, respecto a los rankings definidos por las medidas \mathcal{M}_1 y \mathcal{M}_2 , aquella clase que es mejor representada por la palabra visual K .

Sea $\Theta_{\mathcal{M}}(K, c_j)$ los valores que alcanzan las palabras visuales del conjunto $K = \{k_1, k_2, \dots, k_N\}$ para la clase c_j , de acuerdo a una medida \mathcal{M} dada. Asumamos que estos valores están ordenados descendientemente. Sea $\Phi(k, c_j)$ la posición de la palabra visual $k \in K$ in $\Theta_{\mathcal{M}}(K, c_j)$. Sea $P_k = \min_{c_j \in C} (\Phi(k, c_j))$ la mejor posición de la palabra visual k en el conjunto de todas la clases de objeto $C = \{c_1, c_2, \dots, c_Q\}$. Sea $c_k = \arg \min_{c_j \in C} (\Phi(k, c_j))$ la clase de objetos para la cual k tiene la posición P_k . Luego, la distintividad inter-clase (medida \mathcal{M}_3) de una palabra visual k de acuerdo a una medida \mathcal{M} , se define como sigue:

$$\mathcal{M}_3(k, \mathcal{M}) = \frac{1}{(|C| - 1)(|K| - 1)} \sum_{c_j \neq c_k} (\Phi(k, c_j) - P_k). \quad (4)$$

Métodos para reducir el tamaño del vocabulario visual

En esta sección se presentan dos métodos para rankear y filtrar el vocabulario visual, buscando vocabularios visuales más compactos y confiables.

El primer método, nombrado MMM, está basado en las medidas propuestas en las secciones anteriores. Sean $\Theta^{\mathcal{M}_1}(K)$ y $\Theta^{\mathcal{M}_2}(K)$ los rankings del vocabulario K , usando las medidas $\mathcal{M}_3(K, \mathcal{M}_1)$ y $\mathcal{M}_3(K, \mathcal{M}_2)$, respectivamente. $\Theta^{\mathcal{M}_1}(K)$ y $\Theta^{\mathcal{M}_2}(K)$ brindan un ranking del vocabulario visual basado en el poder distintivo de las palabras visuales, de acuerdo a la variabilidad intra e inter-clase. Con el objetivo de encontrar un consenso, $\Theta(K)$, entre los ranking $\Theta^{\mathcal{M}_1}(K)$ y $\Theta^{\mathcal{M}_2}(K)$ se propone utilizar cualquier método de consenso basado en votos, que esté reportado en la literatura; en nuestro caso, se decidió utilizar el algoritmo Borda Count ([Emerson, 2013](#)), aunque cualquier otro podría haberse usado. El algoritmo Borda Count obtiene un ranking final a partir de múltiples rankings del mismo conjunto. Dadas $|K|$ palabras visuales y un ranking sobre ellas, una palabra visual recibe $|K|$ puntos si está en primer lugar, $|K| - 1$ puntos si está en segundo, $|K| - 2$ si es tercera y así. Para cada palabra, los puntos recibidos por cada ranking son sumados para obtener el ranking final. A partir

de este último ranking, se puede obtener un vocabulario reducido al seleccionar solo las primeras N palabras visuales.

El segundo método de ranking y filtrado, nombrado FRM, está basado en el esquema de pesado *tf.idf*, específicamente, nuestra propuesta se basa en las definiciones introducidas en (Moulin et al., 2010). Tradicionalmente, *tf.idf* ha sido utilizado como esquema de pesado en la representación de las imágenes. No obstante, en nuestra propuesta se utiliza para obtener un ranking y poder filtrar el vocabulario inicial. Sea $D = \{m_1, m_2, \dots, m_N\}$ el conjunto de imágenes de entrenamiento a partir del cual se construye el vocabulario visual. De acuerdo a (Moulin et al., 2010), la *frecuencia del término* y la *frecuencia inversa de documento* de una palabra visual v_i en una imagen m_j , denotadas por tf_{v_i, m_j} y idf_{v_i, m_j} , respectivamente, se definen por las siguientes expresiones:

$$tf_{v_i, m_j} = \frac{K_1 \cdot O_{ij}}{O_{ij} + K_2 \cdot \left(1 - b + b \cdot \left(\frac{|\{v_q | O_{qj} > 0\}|}{V_{avg}}\right)\right)} \quad idf_{v_i, m_j} = \log \frac{|D| - |D_{v_i}| + 0,5}{|D_{v_i}| + 0,5}, \quad (5)$$

donde K_1, K_2 y b son constantes, O_{ij} es la ocurrencia de v_i en la imagen m_j , V_{avg} es el número promedio de palabras visuales que representan a las imágenes del conjunto de entrenamiento, y D_{v_i} es el conjunto de imágenes en las cuales la ocurrencia de v_i es mayor que cero.

Tomando en cuenta la forma en que se construye el histograma de ocurrencias de una palabra visual, es altamente probable que cualquier palabra visual *ocurra* al menos una vez en casi todas las imágenes. Lo anterior puede tener una influencia negativa en el cálculo de la frecuencia inversa de documento. Para solucionar este problema, se propone re-definir el conjunto D_{v_i} como el conjunto de aquellas imágenes en las cuales la ocurrencia de v_i es mayor o igual que la ocurrencia promedio de v_i en las imágenes de entrenamiento. Utilizando las expresión de la frecuencia de término y la nueva definición de la frecuencia inversa de documento, se puede construir para cada palabra visual v_i , un vector que contenga el producto de tf_{v_i, m_j} y idf_{v_i, m_j} , en cada imagen $m_j \in D$. El valor promedio contenido en este vector constituye el ranking v_i . A partir de este ranking se puede obtener un vocabulario reducido si se seleccionan las primeras N palabras visuales del mismo.

Nueva propuesta para la representación de las imágenes

Una vez que se construye el vocabulario visual, las imágenes son representados a través del histograma de ocurrencias de las palabras visuales. Para construir este histograma el poder distintivo y representativo de las palabras visuales no se tiene en cuenta. A continuación, se propone un nuevo método, nombrado SWIR, para construir la representación de las imágenes, que se enfoca en resolver el efecto negativo que causa el problema mencionado anteriormente, sobre el histograma de ocurrencias

Sea $\Theta(N)$ el ranking final de las N palabras que constituyen el vocabulario visual, seleccionadas utilizando uno de los métodos de filtrado propuestos en este trabajo. En lo siguiente se asume que estos valores fueron normalizados tal que están en el intervalo $[0,1]$. Estos valores se utilizarán en la continuación para apoyar la presencia de las palabras altamente distintivas y representativas, así como para penalizar la presencia de aquellas con un bajo poder descriptivo. Para hacer esto, primeramente se determina un elemento *pivote*, denotado por $P_{\Theta(N)}$, como el promedio de valores de ranking existente entre las palabras del vocabulario visual. El *peso de contribución* de una palabra visual v_i , denotado por cw_{v_i} , se calcula de la siguiente forma: $cw_{v_i} = 1 - P_{\Theta(N)} + \Theta_{v_i}$, donde Θ_{v_i} es el valor que tiene v_i en $\Theta(N)$.

Para obtener la representación de una imagen m_j , se propone multiplicar las ocurrencias de las palabras visuales en el histograma de ocurrencias de m_j , por sus respectivos pesos de contribución. De esta forma, aquellas palabras visuales con un ranking superior a $P_{\Theta(N)}$ son consideradas como más representativas y distintivas y consecuentemente, su presencia dentro de la imagen es premiada (i.e., es incrementada). Por otra parte, toda palabra visual con un valor de ranking menor que el pivote es penalizada a través de una reducción de su presencia en el histograma de ocurrencias.

Resultados experimentales

Los experimentos que se presentan en esta sección tienen como objetivo evaluar cuantitativamente las mejoras que nuestras propuestas traen sobre el enfoque BoW tradicional y comparar dichas propuestas con el método de Zhang *et al.* (Zhang *et al.*, 2014), el cual alcanza los mejores resultados de clasificación entre los métodos de selección de descriptores y compresión de la representación de las imágenes, en el contexto de la categorización de objetos. Los experimentos se realizaron en el conjunto Caltech-101 (Fei-Fei *et al.*, 2006), sobre una computadora con un procesador Intel i7 con 3.6 GHz y 64GB RAM.

En los experimentos las imágenes fueron representadas utilizando el esquema BoW, usando los descriptores PHOW e histogramas espaciales como descriptores de las imágenes. Para construir el vocabulario visual se utilizó una variante del algoritmo K-means (Elkan, 2003) con cuatro valores diferentes para K ($K = 512, 1024, 2048$ y 4096); estos vocabularios constituyen las líneas base de comparación. Cada una de las líneas bases fue procesada utilizando el método propuesto en (Zhang *et al.*, 2014), así como los métodos de ranking y filtrado propuestos en este trabajo, MMM y FRM, con y sin utilizar el método de representación propuesto, SWIR. Posteriormente, nueve nuevos vocabularios fueron obtenidos seleccionando de las líneas bases 10 %, 20 %, ..., 90 %, de las palabras visuales con mejor ranking, respectivamente. Para definir el conjunto de entrenamiento y prueba se siguió el protocolo propuesto en (Lazebnik *et al.*, 2006); esto es, para el conjunto de entrenamiento se tomaron aleatoriamente 30 imágenes por cada clase de objeto y para el conjunto de entrenamiento se tomaron

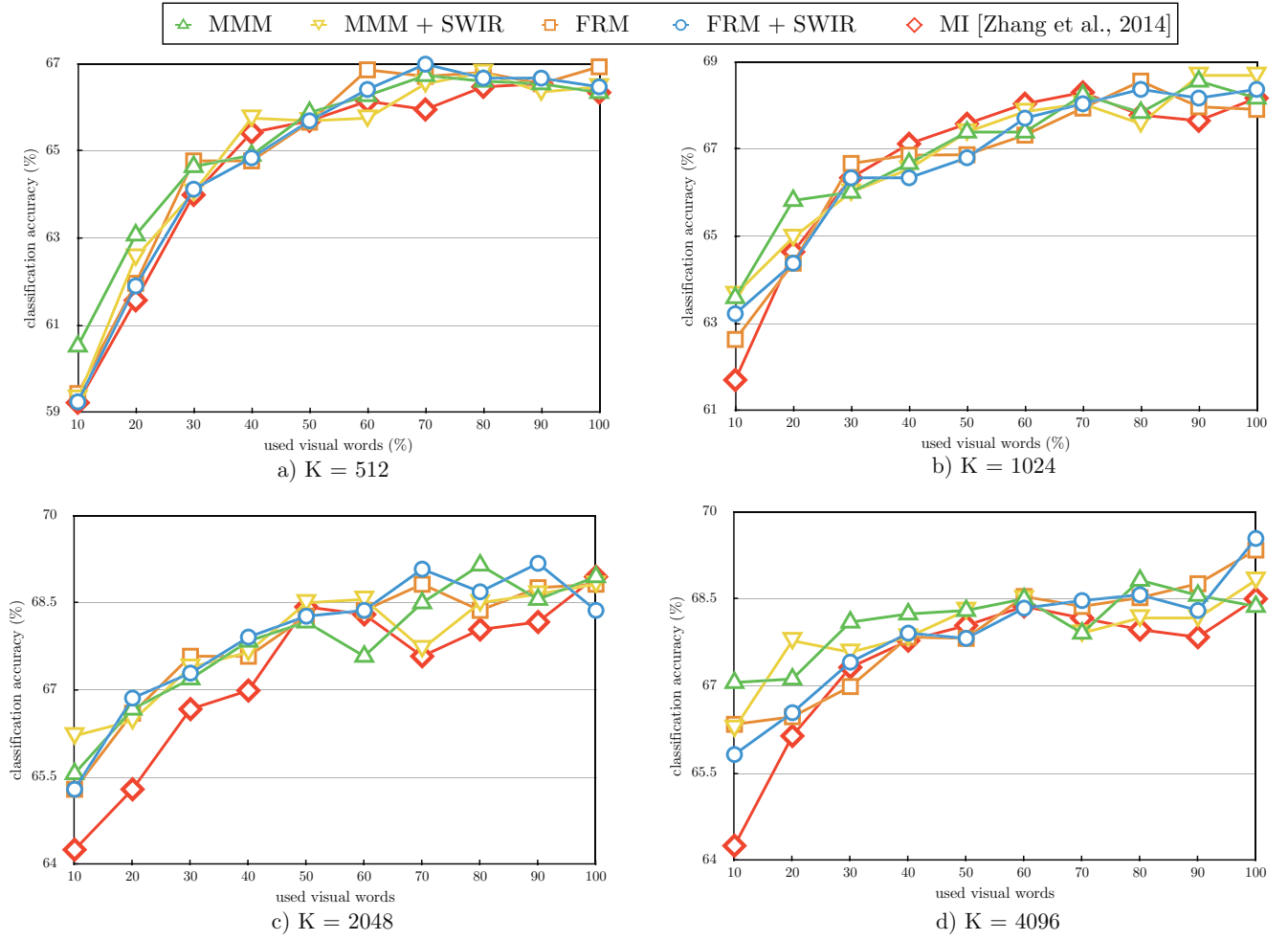


Figura 1. Eficacia obtenida por cada método en la clasificación del conjunto Caltech-101.

el resto de las imágenes de cada clase, limitadas por 50. Los vocabularios obtenidos fueron utilizados en la tarea de clasificación de objetos, empleando una optimización de SVM propuesta en (Vedaldi and Zisserman, 2011). La eficacia obtenida por cada método se muestra en la Figura 1.

Como puede notarse en la Figura 1, para cada valor de K usado en el experimento, nuestras propuestas alcanzan los mejores resultados de clasificación para los valores más altos de compresión del vocabulario (10 y 20 %), siendo MMM el mejor método en este caso. Para el resto de los tamaños de filtrado nuestras propuestas obtienen resultados comparables e incluso superiores en varios casos, que el método de Zhang *et al.* Zhang et al. (2014). Adicionalmente, en casi todos los valores de K , la combinación de los métodos MMM y FRM, aplicando el método SWIR permite obtener los resultados más altos de eficacia. Por lo tanto, podemos afirmar que tener

en cuenta la distintividad y la representatividad de las palabras visuales en el proceso de representación de las imágenes, permite aumentar la eficacia de los clasificadores.

Conclusiones

En este trabajo se han propuesto tres propiedades que permiten evaluar la habilidad de una palabra para representar y distinguir a una clase de objeto. Adicionalmente, se propusieron medidas que permiten evaluar cuantitativamente el grado de cumplimiento de las medidas comentadas, por parte de una palabra visual. También se propusieron dos métodos de ranking y filtrado de vocabulario visual, que permiten obtener vocabularios más distintivo y representativo. Por último, también se introdujo un nuevo método para representar las imágenes a partir del vocabulario visual, teniendo en cuenta el poder distintivo de las palabras de dicho vocabulario.

Los experimentos llevados a cabo sobre el conjunto Caltech-101 mostraron que usando palabras visuales más discriminativas y representativas es muy posible obtener vocabularios más compactos y eficaces. Estos experimentos también mostraron que si se tiene en cuenta el valor descriptivo y distintivo de las palabras visuales en el momento en que se construye el histograma de ocurrencias de dicha imagen se mejora la representación de las imágenes.

El trabajo futuro se centrará en definir un método que permita determinar de manera automática el corte en el vocabulario visual que maximiza la eficacia del clasificador.

Referencias

- Leonardo Chang, Miriam M. Duarte, L. E. Sucar, and Eduardo F. Morales. A bayesian approach for object classification based on clusters of sift local features. *Expert Systems with Applications.*, 39:1679–1686, 2012.
- Charles Elkan. Using the triangle inequality to accelerate k-means. In Tom Fawcett and Nina Mishra, editors, *ICML*, pages 147–153, 2003.
- Peter Emerson. The original borda count and partial voting. *Social Choice and Welfare*, 40(2):353–358, 2013.
- Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006. URL <http://www.ncbi.nlm.nih.gov/pubmed/16566508>.
- Y. Gong, S. Kumar, H. A. Rowley, and S. Lazebnik. Learning binary codes for high-dimensional data using bilinear projections. In *CVPR 2013*, 2013.

- H. Jégou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *IEEE Pattern Analysis and Machine Intelligence*, 33(1):117–128, 2011.
- Yu-Gang Jiang, Chong-Wah Ngo, and Jun Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, CIVR '07, pages 494–501, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-733-9. doi: 10.1145/1282280.1282352. URL <http://doi.acm.org/10.1145/1282280.1282352>.
- Mingyuan Jiu, Christian Wolf, Christophe Garcia, and Atilla Baskurt. Supervised learning and codebook optimization for bag of words models. *Cognitive Computation*, 4:409–419, December 2012.
- Kraisak Kesorn and Stefan Poslad. An enhanced bag-of-visual word vector space model to represent visual content in athletics images. *IEEE Transactions on Multimedia*, 14(1):211–222, 2012.
- Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, pages 2169–2178, 2006.
- R.J. Lopez-Sastre, T. Tuytelaars, F.J. Acevedo-Rodriguez, and S. Maldonado-Bascon. Towards a more discriminative and semantic visual vocabulary. *Computer Vision and Image Understanding*, 115(3):415 – 425, 2011. ISSN 1077-3142. Special issue on Feature-Oriented Image and Video Computing for Extracting Contexts and Semantics.
- David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. ISSN 0920-5691.
- C. Moulin, C. Barat, and C. Ducottet. Fusion of tf.idf weighted bag of visual features for image classification. In *Content-Based Multimedia Indexing (CBMI), 2010 International Workshop on*, pages 1–6, June 2010. doi: 10.1109/CBMI.2010.5529901.
- A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *Pattern Analysis and Machine Intelligence*, 34(3), 2011.
- Y. Zhang, J. Wu, and J. Cai. Compact representation for image classification: To choose or to compress? In *CVPR 2014*, pages 907–914, 2014.