Intersecciones Semánticas entre Visión Artificial y Mirada Artística

Pilar Rosado[1], Eva Figueras[2] & Ferran Reverter[3]

1,2,3) University of Barcelona. Spain

Date of publication: February 3th, 2014
Edition period: October 2013-February 2014

To cite this article: Rosado, P., Figueras, E.& Reverter, F. (2014). Intersecciones semánticas entre visión artificial y mirada artística. *Barcelona, Research, Art, Creation*, *Vol 2*(1), 1-54. doi:10.4471/brac.2014.01

To link this article: http://dx.doi.org/10.4471/brac.2014.01

BRAC - Barcelona Research Art Creation. Vol. 2 No. 1, February 2014, pp. 1-54

Semantic Intersections between Artificial and Artistic Vision

Pilar Rosado
University of Barcelona

Eva Figueras
University of Barcelona

Ferran Reverter
University of Barcelona

(Received: 13 November 2013; Accepted: 2 January 2014; Published: 3 February 2014)

Abstract

In our project we have approached the difficulties of automatic classification of images on which the conception and design of sculptor M. Planas artistic production are based. This artist constantly generates images in his creative process. The methodology used is based on local characteristics. In order to build up a visual vocabulary for basing image description on, we followed a procedure similar to the one used in automatic text analysis. The method is known as the "Bag-of-Words" (BOW) model because every document is represented as a distribution of frequencies of the words in the text, without considering the syntactic relationships among them. In the sphere of images we refer to "Bag-of-Visual Terms" (BOV) representations. This approach consists in analysing images as a group of regions, describing only their appearance without taking into account their spatial structure. To overcome the disadvantages of polysemy and synonymy that this methodology has associated, we have used probabilistic latent semantic analysis (PLSA), that detects underlying topic in images. The outcomes are promising, the described cataloguing method may provide the artist with new viewpoints for future works.

Keywords: Artificial vision, Bag-of-visterms, SIFT descriptors, image cataloging, automated image analysis, probabilistic latent semantic analysis.

ISSN: 2014-8992
DOI: 10.4471/ brac.2014.01

BRAC - Barcelona Research Art Creation. Vol. 2 No. 1, February 2014, pp. 1-54

Intersecciones Semánticas entre Visión Artificial y Mirada Artística

Pilar Rosado
Universidad de Barcelona

Eva Figueras
Universidad de Barcelona

Ferran Reverter
Universidad de Barcelona

(*Recibido: 13 Noviembre 2013; Aceptado: 2 Enero 2013; Publicado: 3 Febrero 2014)*

Resumen

En el presente artículo se ha desarrollado un sistema capaz de categorizar de forma automática la base de datos de imágenes que sirven de punto de partida para la ideación y diseño en la producción artística del escultor M. Planas. La metodología utilizada está basada en características locales. Para la construcción de un vocabulario visual se sigue un procedimiento análogo al que se utiliza en el análisis automático de textos (modelo "Bag-of-Words"-BOW) y en el ámbito de las imágenes nos referiremos a representaciones "Bag-of-Visual Terms" (BOV). En este enfoque se analizan las imágenes como un conjunto de regiones, describiendo solamente su apariencia e ignorando su estructura espacial. Para superar los inconvenientes de polisemia y sinonimia que lleva asociados esta metodología, se utiliza el análisis probabilístico de aspectos latentes (PLSA) que detecta aspectos subyacentes en las imágenes, patrones formales. Los resultados obtenidos son prometedores y, además de la utilidad intrínseca de la categorización automática de imágenes, este método puede proporcionar al artista un punto de vista auxiliar muy interesante.

Palabras clave: Visión artificial, Bag-of-visterms, descriptores SIFT, catalogación de imágenes, análisis automático de imágenes, análisis probabilístico de aspectos latentes

ISSN: 2014-8992
DOI: 10.4471/ brac.2014.01

Se ha dejado adormecer nuestra capacidad innata de entender con los ojos, y hay que volver a despertarla" (Arnheim, 1983, p.13). Quizá una forma de conseguirlo sea con la ayuda de los métodos de visión artificial que tenemos a nuestro alcance en la actualidad, y a ello pretendemos contribuir con este trabajo.

Muchas veces resulta difícil expresar con palabras lo que se percibe en una imagen. Es normal que podamos ver y sentir las cualidades de una obra sin poder explicarlas. El problema no sólo reside en el lenguaje, sino en que aún no hemos sido capaces de plasmar esas cualidades percibidas en las categorías adecuadas. Para poder nombrar algo debemos haberlo visto, oído, pensado o sentido con anterioridad (Arnheim, 1983, p.15). A estas consideraciones se suma el hecho de que la mirada hacia el mundo es un juego de equilibrio entre las propiedades del objeto observado y la naturaleza del sujeto que observa.

El artista visual hace uso de sus categorías formales para capturar desde lo particular aquello universalmente significativo, desde una forma necesariamente personal. Muchos artistas utilizan la fotografía como herramienta de proyectación y éste sería el caso del escultor M. Planas (Planas, 2014). Las imágenes que utiliza conforman una parte esencial de su proceso creativo, las plantea como un gran fondo documental sobre el que trabajar posteriormente, de manera especial en el campo escultórico. Las imágenes que el artista captura pueden formar parte, tanto de las fases iniciales de su proceso creativo, y también convertirse en el resultado de su trayecto artístico; pueden ser tanto la herramienta de trabajo como el producto de éste.

El objetivo que nos proponemos en el presente estudio es analizar la base de datos de imágenes que ha generado el artista mediante metodologías de visión artificial e intentar dilucidar si éstas generan de forma automática categorías distinguibles en este conjunto, si son significativas para el artista y si guardan algún tipo de relación con otras obras producidas por él, y así poder determinar si el método descrito es capaz de detectar contenidos semánticos en la colección de imágenes.

Al poner de manifiesto estas categorías visuales resulta más fácil apreciar las relaciones estructurales latentes en las imágenes. No se intenta

substituir la intuición espontánea o el criterio experto, sino agudizar, reforzar o evidenciar nuevos elementos de utilidad para la mirada (o percepción visual). La categorización automática de imágenes, además de su utilidad intrínseca, puede proporcionar al artista un punto de vista auxiliar muy interesante.

Para realizar este estudio ha sido fundamental tener al alcance un fondo consolidado de aproximadamente 3.000 imágenes digitales, todas ellas pertenecientes al mismo autor y con un perfil coherente, pero de estructura y características formales, conceptuales y temáticas diversas. La participación e implicación directa en este proceso del propio autor también ha permitido contrastar los resultados.

Al proponer una investigación de estas características dentro del ámbito de las Bellas Artes se planteó que los resultados obtenidos se podrían extrapolar a todo tipo de acciones entorno a la creación, en las que la comparación entre imágenes fuera su característica principal, logrando aplicaciones encaminadas al aprendizaje, al conocimiento y a la investigación en imágenes.

Metodología

La representación de la imagen digital es un elemento clave para su clasificación, anotación, segmentación o recuperación. Casi todos los métodos de visión por computador, cuando se enfrentan al problema del análisis del contenido de una imagen, recurren a funciones adecuadas para describirlo de forma compacta. Este sería el caso de los procedimientos basados en características locales que producen una representación de la imagen versátil y sólida capaz de mostrar el contenido global y local al mismo tiempo, y a la vez hacen robusta la descripción ante la oclusión parcial de objetos contenidos y la transformación de la propia imagen.

Representación de la imagen

Para la construcción de un vocabulario visual en el que basar la descripción de las imágenes, seguimos un procedimiento análogo al que se utiliza en el análisis automático de textos (Joachims, 1998). Se conoce como modelo "Bag-of-Words" (BOW) porque cada documento está representado como

una distribución de frecuencias de las palabras presentes en el texto, sin tener en cuenta las relaciones sintácticas existentes entre ellas.

En el ámbito de las imágenes nos referiremos a representaciones "Bag-of-Visual Terms" (BOV). Este enfoque consiste en analizar las imágenes como un conjunto de regiones, describiendo solamente su apariencia e ignorando su estructura espacial. La representación BOV se construye a partir de la extracción y cuantización automática de descriptores locales y ha demostrado ser una de las mejores técnicas para resolver diferentes tareas en la visión por computador. La representación BOV fue implementada por primera vez (Willamowski, Arregui, Csurka, Dance & Fan, 2004) en el desarrollo de una sistema experto de reconocimiento de objetos.

La construcción BOV requiere dos decisiones principales de diseño:

a. La elección de los descriptores locales que aplicamos en nuestras imágenes (Ver anexo A).
b. La elección del método que se utilice para obtener el vocabulario visual (Ver anexo B).

Ambas decisiones pueden influir en el rendimiento del sistema resultante, sin embargo la representación BOV es robusta, conserva su buen comportamiento en un amplio rango de opciones de los parámetros.

La Fig. 1 resume el proceso para obtener la representación BOV de las imágenes de una colección.

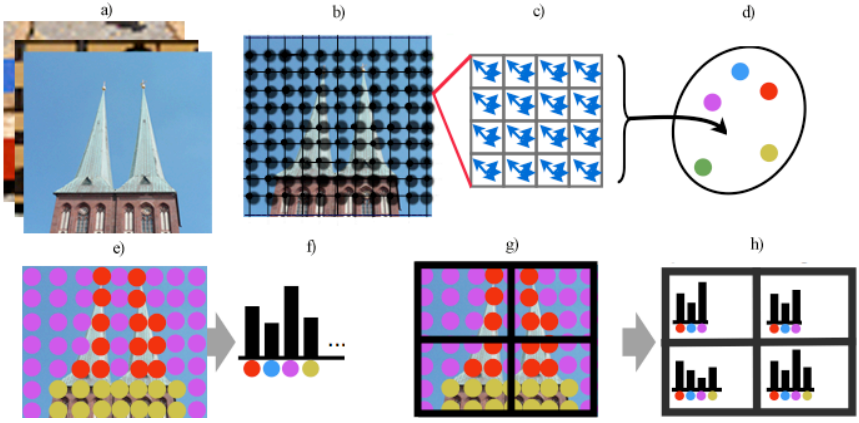


Figura 1. a) Colección de imágenes. b) Se localizan los nodos en una malla regular. c) Se calculan los descriptores SIFT de los nodos. d) Se cuantizan los descriptores de todos los nodos en M clústeres, los cuales definirán un vocabulario visual de M

palabras visuales. e) Una vez se dispone del vocabulario, los descriptores de cada imagen se asignan a la palabra visual más cercana. f) Para obtener la representación BOV de una imagen dada, se calcula la frecuencia de cada palabra visual en la imagen. g) Secuencia de cuadrículas sobre la imagen para elaborar los histogramas en pirámide h) para así tener en cuenta la relación espacial entre palabras visuales.

Esta representación de una imagen no contiene información acerca de las relaciones espaciales entre palabras visuales, del mismo modo que la representación BOW no tiene en cuenta la información relativa al orden de las palabras en los documentos (Fig. 2).

No obstante, los métodos BOV, que representan una imagen como una colección desordenada de características locales, han demostrado impresionantes niveles de rendimiento en tareas de categorización de imágenes completas. Sin embargo, debido a que estos métodos no tienen en cuenta toda la información acerca de la disposición espacial de las características, se ha visto limitada su capacidad descriptiva. En particular, son incapaces de capturar formas o de separar un objeto de su fondo.

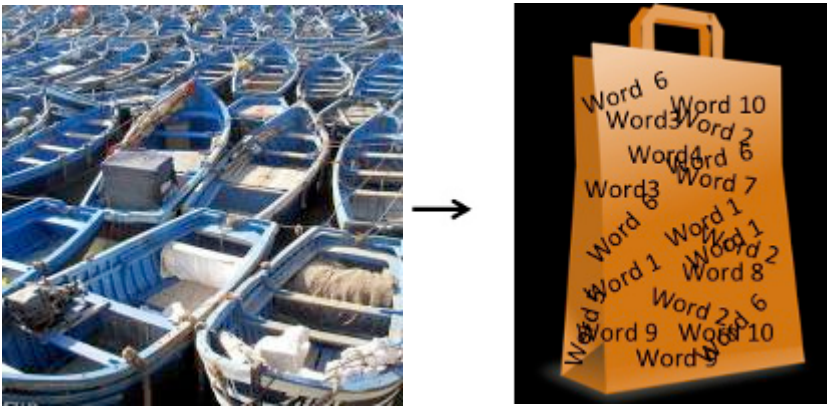


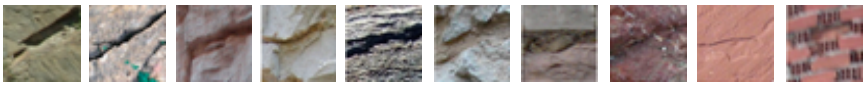
Figura 2. La representación BOV de una imagen no contiene información acerca de las relaciones espaciales entre palabras visuales que la componen.

Para superar las limitaciones del enfoque BOV hemos implementado una metodología de histogramas en pirámide que configura una secuencia cada vez más fina de cuadrículas sobre la imagen y lleva a cabo un análisis tipo BOV en cada una de las cuadrículas, obteniendo finalmente una suma ponderada de la cantidad de coincidencias que ocurren en cada nivel de resolución de la pirámide (Grauman, K. & Darrel, T., 2005).

Representación de aspectos latentes

La representación BOV es fácil de construir. Sin embargo, adolece de dos inconvenientes (Fig. 3): polisemia (una sola palabra visual puede representar diferentes contenidos de la escena) y sinonimia (varias palabras visuales pueden caracterizar el mismo contenido de la imagen).

Palabra 1



Palabra 2

Figura 3. Muestras de regiones de imágenes correspondientes a dos palabras visuales de un vocabulario de 300 palabras. Podemos considerar que ambas palabras describen un contenido común; textura rocosa no homogénea, y en este sentido representan una sinonimia. Además, vemos que dentro de una misma palabra hay regiones que representan contenidos distintos, en unos casos el contenido es roca y en otros es muro, por tanto podríamos considerarlas polisémicas.

Para solventar en parte los inconvenientes anteriores, utilizaremos el análisis probabilístico de aspectos latentes (PLSA), una metodología original de la minería de textos (Hofmann, 2001).

Las aplicaciones del PLSA en el análisis estadístico de textos están orientadas a descubrir automáticamente los temas tratados en un

documento, tomando como punto de partida la representación BOW de documentos.

La extensión del PLSA hacia el análisis de imágenes pasa por considerar las imágenes como documentos con un vocabulario visual establecido a partir de un proceso de cuantización como se señala en el anexo B.

El PLSA detectará en las imágenes categorías de objetos, patrones formales, de modo que una imagen que contiene varias tipos de objetos se modela como una mezcla de temas (Fig. 4).

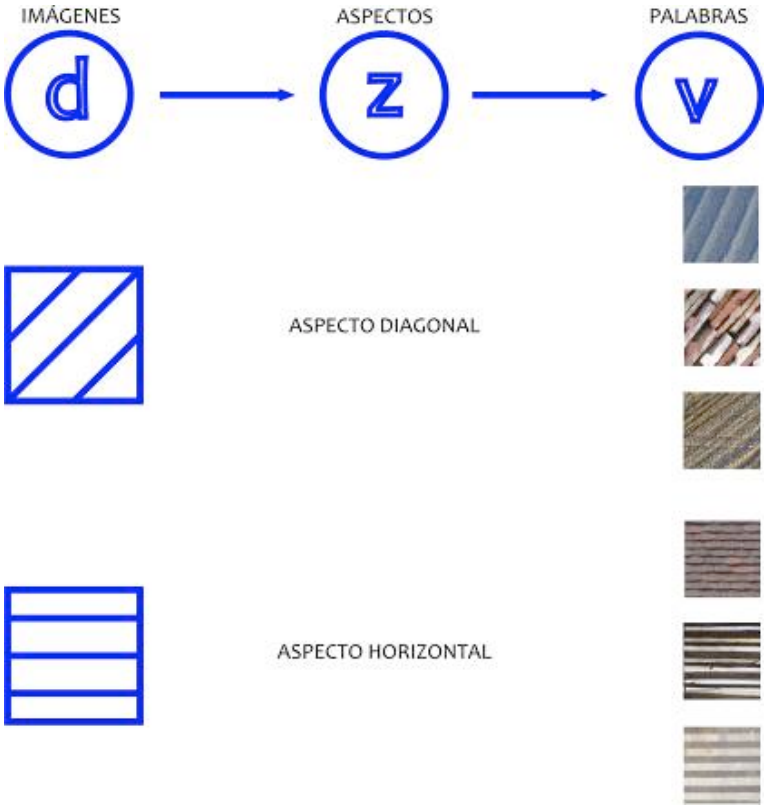


Figura 4. El método PLSA captura la co-ocurrencia de palabras visuales entre imágenes.

El método PLSA está descrito de forma más exhaustiva en el anexo C.

A partir de esta metodología nos ha sido posible analizar la colección de imágenes y encontrar aspectos subyacentes mediante los cuales catalogar toda la colección y también contrastar dichos aspectos o patrones con los propuestos por el autor.

Implementación

El desarrollo se ha llevado a cabo mediante scripts escritos en MATLAB, versión 2013a (The MathWorks) (8.1.0.604). Los descriptores SIFT y el vocabulario de palabras visuales se han implementado mediante funciones disponibles en la biblioteca de código abierto VLFeat, versión 0.9.16 (Vedaldi & Fulkerson, 2008). El PLSA ha sido implementado mediante funciones desarrolladas por los propios autores.

Resultados

La muestra inicial de nuestro estudio está compuesta por un total de 2.846 imágenes fotográficas capturadas por el propio artista. Se trata de un conjunto de imágenes tomadas, la mayoría, de exteriores y captadas desde diferentes ángulos y detalles (llegando a fragmentos y particularidades que se pueden captar como elementos abstractos y/o texturados). El tamaño de las imágenes del artista está entre 480 x 480 píxeles y 1400 x 1400 píxeles, pero el proceso reescala a 480 píxeles las imágenes que superan este tamaño para garantizar un tratamiento homogéneo de la colección de fotografías.

Con el total de imágenes se han realizado diversas pruebas del sistema fijando el número de aspectos a encontrar en 20, 15, 10 y finalmente 5. Los resultados obtenidos con 5 aspectos mostraban que algunas tipologías no resultaban visibles, y con 20 y 15 aspectos las categorías quedaban poco representadas y algo dispersas. Por lo tanto, se decidió que la prueba más representativa para el total de imágenes analizadas era la que clasificaba la muestra en 10 aspectos.

Para la evaluación del resultado de esta agrupación se consideran prioritariamente imágenes tipificadas en un determinado aspecto con una probabilidad igual o superior a 0.6. La forma en que se calcula dicha

probabilidad está detallada en el anexo C de este trabajo. A pesar de esta consideración, en la discusión de cada aspecto se muestran las 3 imágenes con mayor probabilidad de estar asociadas a ese aspecto porque se considera que así el lector puede hacerse más fácilmente a la idea de la consistencia visual del mismo. Los casos que tienen alguna característica conflictiva o dudosa se comentan de forma más concreta para facilitar la comprensión.

Recordamos que la evaluación computacional se efectúa en escala de grises.

Ha resultado complicado asignar un descriptor textual a los aspectos hallados por el sistema, ya que no siempre es fácil y directo asociar el contenido visual del conjunto de imágenes de un aspecto con una descripción literal. Los aspectos hallados por la máquina no sólo deben entenderse compositivamente sino que se debe tener presente que se basan en encontrar co-ocurrencias de palabras visuales. A modo de ejemplo aclaratorio, este método sería capaz de detectar y agrupar en el mismo aspecto las imágenes que contengan un rostro, por la co-ocurrencia de las palabras visuales ojos, nariz, boca, etc.

A raíz del primer análisis de los resultados considerando 10 aspectos percibimos que la muestra consta de dos tipologías de imágenes muy marcadas; un tipo de fotografías que presenta un único aspecto muy destacado (las llamaremos imágenes de menos entropía) que son las que aparecen representadas en estas 10 primeras categorías, y otro que presenta varios aspectos asociados simultáneamente (las llamaremos imágenes de más entropía**)** y que no resulta visible en este primer análisis.

Para poder distinguir y tratar separadamente estas dos tipologías hemos utilizado el índice de entropía de Shannon (Cover & Thomas, 2006).

La metodología PLSA (véase anexo C) proporciona una distribución de probabilidad de los aspectos en las imágenes. Esto es, para una imagen dada *d*, tenemos un vector de probabilidades:

$$\left(p(z_1 / d),.p(z_2 / d),...,p(z_K / d)\right)$$

De donde, podemos calcular el índice de Entropía de Shannon de la imagen *d*, mediante:

$$H(d) = -\sum_{i=1}^{K} p(z_i / d)\log(p(z_i / d)$$

De esta manera una imagen que esté asociada a un único aspecto, es decir, una imagen con vector de probabilidades con todo ceros excepto un uno, su valor de entropía será mínimo e igual a H(d)=0, contrariamente, una imagen que esté asociada por igual a todos los aspectos, es decir, con vector de probabilidades con 1/10 en cada componente, su valor de entropía será máximo e igual a H(d)=2.3026.

Los rangos de entropía teóricos respecto a 10 aspectos irían de 0 a 2.3026. Los observados en nuestra muestra van de prácticamente 0 a 2,17. Las imágenes que tienen una entropía elevada son aquellas que el procedimiento ha asociado de manera equiprobable a cada uno de los aspectos.

De esta forma se decide seleccionar del total de la muestra las imágenes con un valor de entropía superior a 1,4 y repetir de nuevo la búsqueda de aspectos en este nuevo conjunto formado por 1.482 imágenes. Se repite de nuevo todo el proceso generando los descriptores locales, el vocabulario visual y se intenta así que el sistema sea capaz de establecer nuevas relaciones entre imágenes visualmente más complejas dando lugar a nuevos aspectos latentes distintos de las 10 primeros. La prueba resulta un éxito y se generan otro conjunto distinto de 10 aspectos sobre la nueva muestra. En total el sistema es capaz de categorizar en 20 grupos el total de imágenes analizadas y estos son los resultados que pasaremos a discutir en el resto del apartado.

A continuación se muestra una selección de las categorías de imágenes menos entrópicas y más entrópicas acompañadas por los respectivos histogramas de aspectos, numerados del 1 al 10 en el eje de abscisas y en el eje de ordenadas se indica la probabilidad asociada a cada aspecto.

En el caso de las imágenes de la categoría menos entrópica, se observa que la probabilidad se concentra mayoritariamente en un aspecto (Fig. 5).

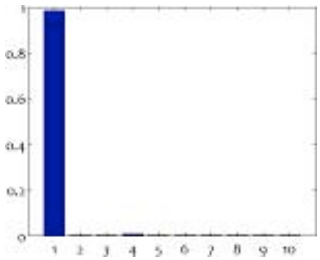


Figura 5. Histograma de una imagen de la categoría menos entrópica.

En el caso de las imágenes de la categoría más entrópica la distribución de probabilidad se reparte entre diversos aspectos (Fig. 6).

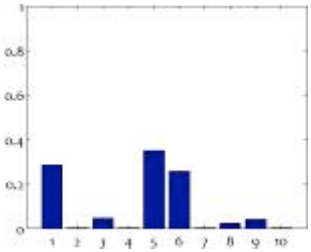


Figura 6. Histograma de una imagen de la categoría más entrópica.

Imágenes e histogramas de las imágenes menos entrópicas:

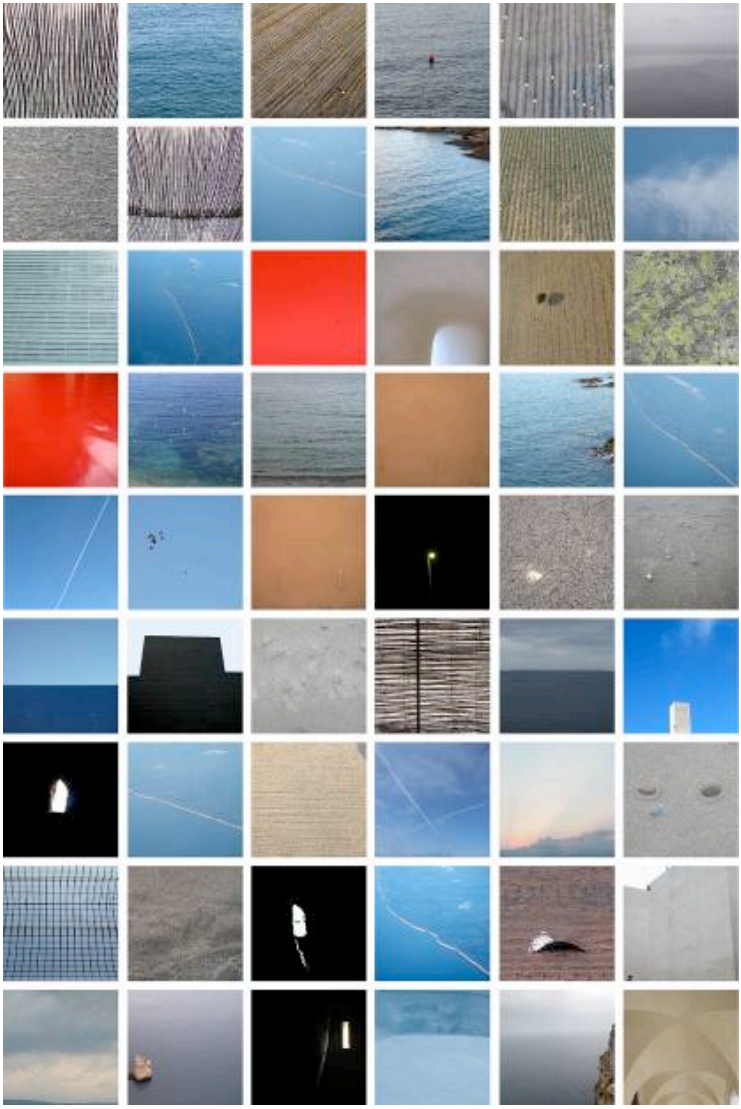


Figura 7. Conjunto de imágenes de la categoría menos entrópica.

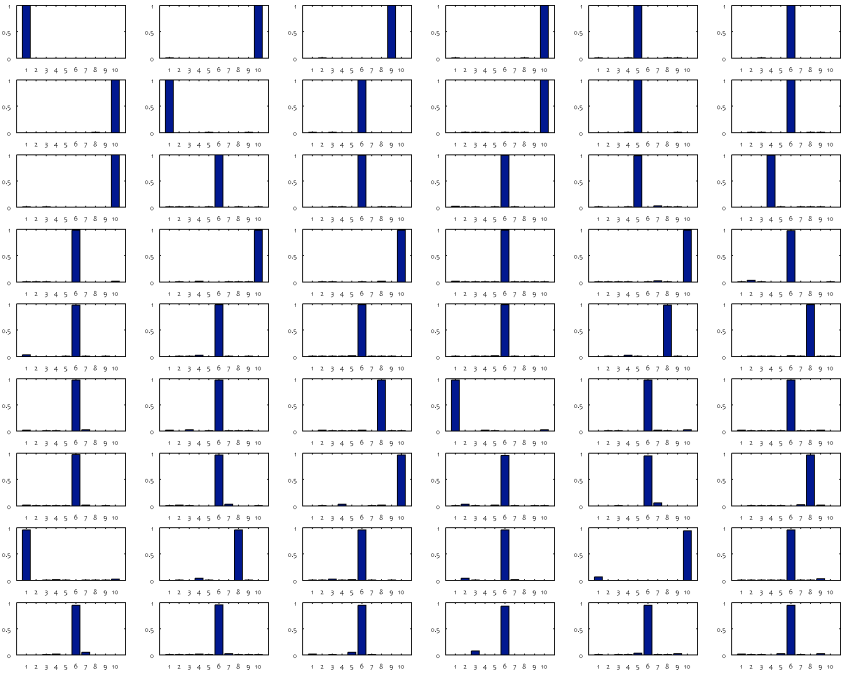


Figura 8. Conjunto de histogramas de las imágenes de la categoría menos entrópica.

Imágenes e histogramas de las imágenes más entrópicas:

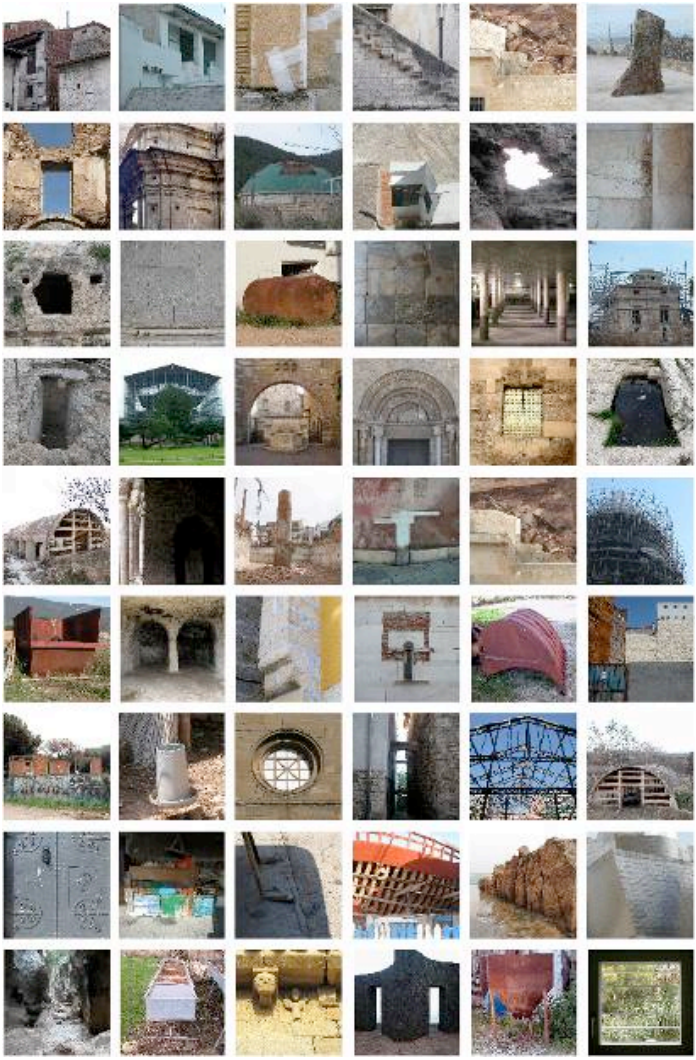


Figura 9. Conjunto de imágenes de la categoría más entrópica.

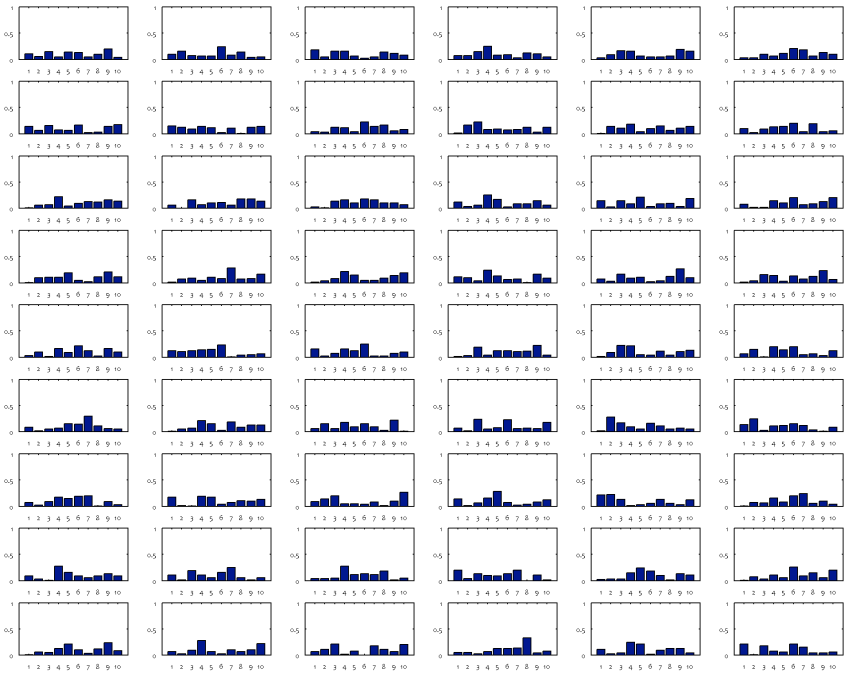


Figura 10. Conjunto de histogramas de las imágenes de la categoría más entrópica.

Aspectos de imágenes menos entrópicas

A partir de ahora los aspectos pertenecientes a esta categoría de imágenes menos entrópicas los nombraremos abreviados como LE (Low entropy).

1. Aspecto-LE1: Líneas Finas Definidas. El aspecto-LE1 Líneas Finas Definidas agrupa un conjunto de imágenes que contienen trazos lineales en las que predomina una dirección, que puede ser horizontal, vertical o ambos (Fig. 11).

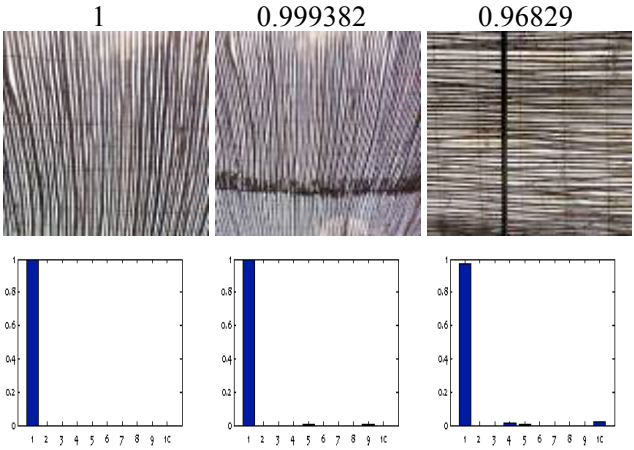


Figura 11. Imágenes tipificadas en el Aspecto-LE1 con mayor probabilidad y los histogramas de aspectos correspondientes.

Vemos que en el histograma de la tercera imagen aumenta ligeramente la probabilidad del aspecto-LE10 Horizontal Vibrante. Para clarificar un poco este aspecto y determinar la importancia de la verticalidad hemos girado al tercera imagen 90 º a la izquierda y hemos vuelto a comprobar el aspecto mayoritario (Fig. 12). El resultado se muestra a continuación:

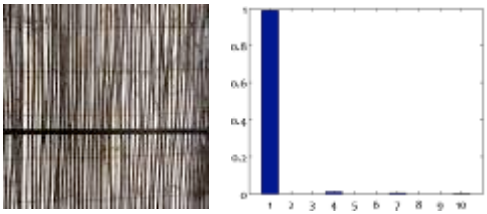


Figura 12. Tercera imagen de la Fig. 11 girada 90 º a la izquierda y el histograma de aspectos correspondiente.

Se puede concluir que la verticalidad no es determinante para este aspecto.

2. Aspecto-LE2: Diagonal Descendente. El aspecto-LE2 Diagonal Descendente agrupa las imágenes en las que predomina básicamente una direccionalidad de bajada desde el ángulo superior izquierdo al ángulo inferior derecho (Fig. 13).

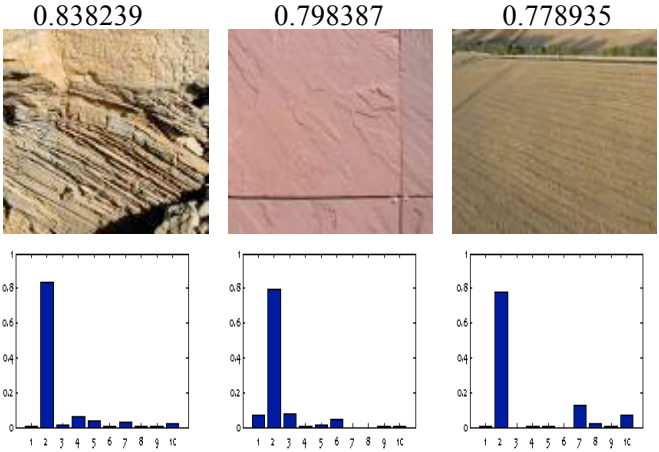


Figura 13. Imágenes tipificadas en el Aspecto-LE2 con mayor probabilidad y los histogramas de aspectos correspondientes.

También observamos en la primera imagen componentes del aspecto-LE4 Textura Heterogénea, que se corresponderían con la zona superior de la imagen. En la segunda imagen vemos una componente perteneciente al aspecto-LE1 Líneas Finas Definidas y en la tercera imagen componentes de los aspectos-LE7 Horizontal Estrecha y aspecto-LE10 Horizontal Vibrante.

3. Aspecto-LE3: Horizontal Amplia. El aspecto-LE3 Horizontal Amplia reúne un conjunto de imágenes caracterizadas por presentar una bandas horizontales anchas muy marcadas (Fig. 14).

Se puede destacar el componente de aspecto-LE10 Horizontal Vibrante de la primera imagen. En la segunda imagen las componentes de aspecto-LE1 Líneas Finas Definidas y aspecto-LE9 Diagonal Ascendente y el aspecto-LE1 Líneas Definidas de la tercera imagen.

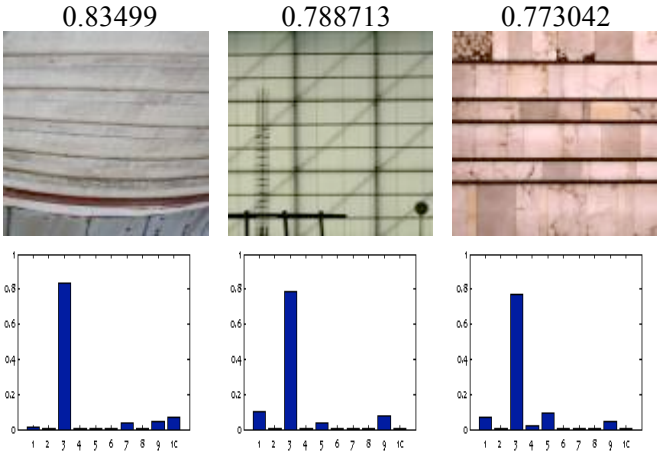


Figura 14. Imágenes tipificadas en el Aspecto-LE3 con mayor probabilidad y los histogramas de aspectos correspondientes.

4. Aspecto-LE4: Textura Heterogénea. El aspecto-LE4 Textura Heterogénea presenta una agrupación bastante sólida, sobre todo teniendo en cuenta la percepción en escala de grises de la computadora (Fig. 15).

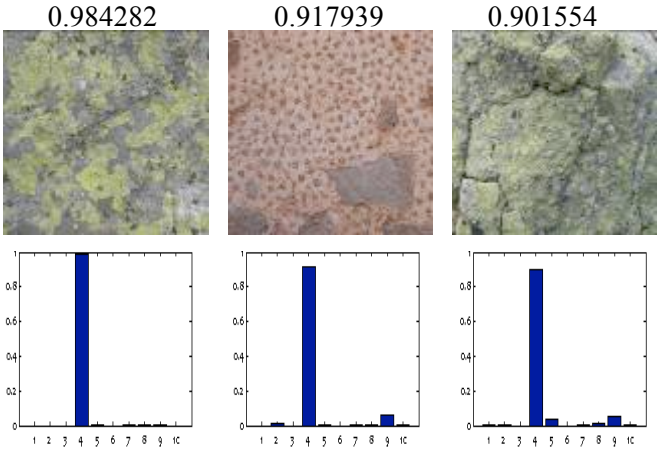


Figura 15. Imágenes tipificadas en el Aspecto-LE4 con mayor probabilidad y los histogramas de aspectos correspondientes.

5. Aspecto-LE5: Vertical Irregular Texturada. El aspecto-LE5 Vertical Irregular Texturada es un aspecto muy compacto y fácilmente identificable (Fig. 16).

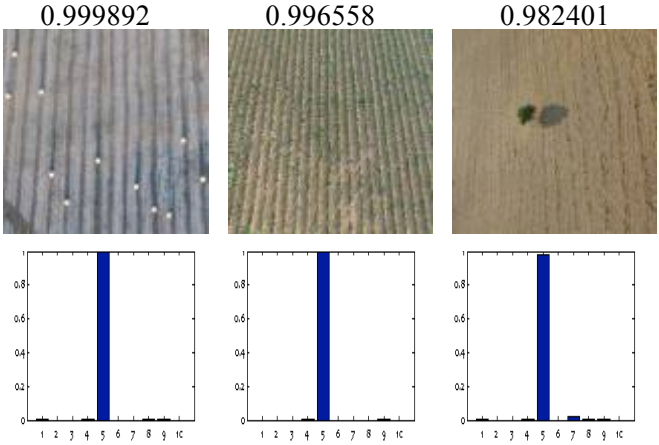


Figura 16. Imágenes tipificadas en el Aspecto-LE5 con mayor probabilidad y los histogramas de aspectos correspondientes.

6. Aspecto-LE6: Liso. Aspecto-LE6 Liso, carente de textura. Para valorar este aspecto en su justa medida hemos de recordar que la máquina sólo considera la escala de grises (Fig. 17).

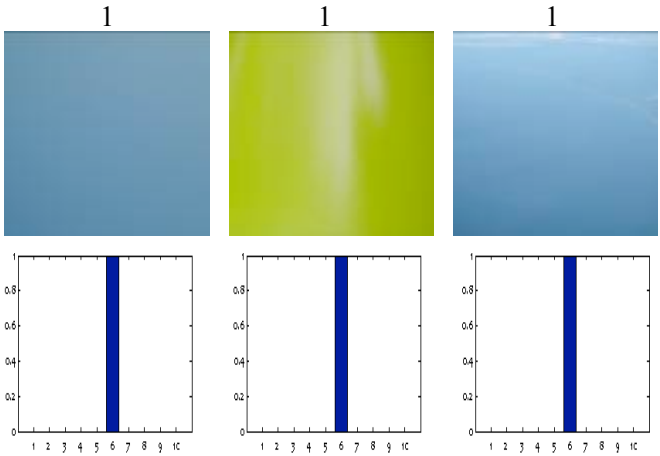


Figura 17. Imágenes tipificadas en el Aspecto-LE6 con mayor probabilidad y los histogramas de aspectos correspondientes.

7. Aspecto-LE7: Horizontal Estrecha. El aspecto-LE7 Horizontal Estrecha presenta unas bandas horizontales más comprimidas que en el aspecto-LE3 Horizontal Amplia descrito con anterioridad (Fig. 18).

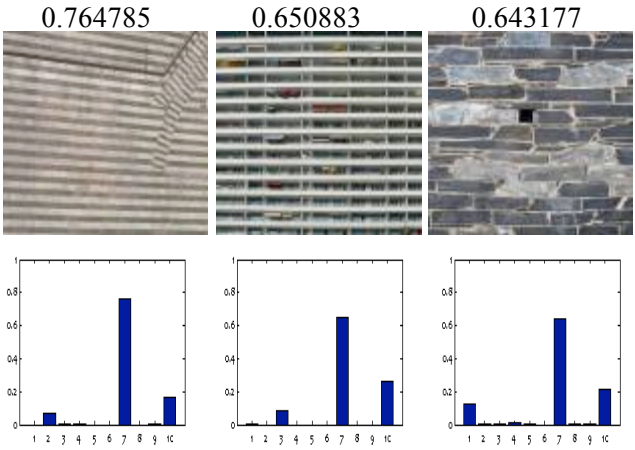


Figura 18. Imágenes tipificadas en el Aspecto-LE7 con mayor probabilidad y los histogramas de aspectos correspondientes.

Cabe destacar la importante componente en las 3 imágenes del aspecto-LE10 Horizontal Vibrante, que por el contrario no está presente en las imágenes del aspecto-LE3 Horizontal Amplia, ya que aquellas presentan líneas horizontales muy bien definidas.

8. Aspecto-LE8: Textura Homogénea. El aspecto-LE8 Textura Homogénea es también compacto y definido, parecido al aspecto-LE6 Liso aunque es ligeramente rugoso. Observamos pequeños elementos que son los responsables de la disminución de la probabilidad (Fig. 19).

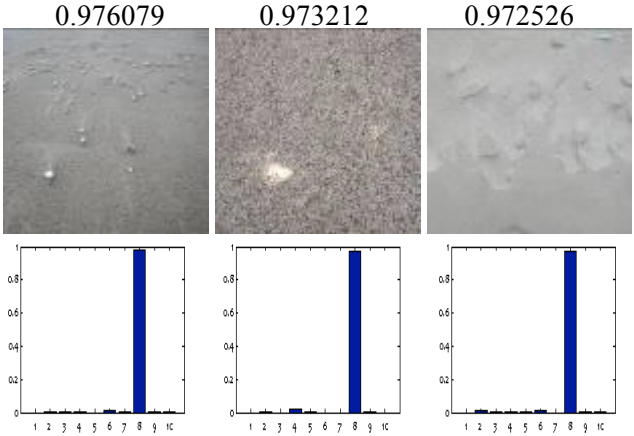


Figura 19. Imágenes tipificadas en el Aspecto-LE8 con mayor probabilidad y los histogramas de aspectos correspondientes.

9. Aspecto-LE9: Diagonal Ascendente. El aspecto-LE9 Diagonal Ascendente agrupa las imágenes en las que predomina básicamente una direccionalidad de subida desde el ángulo inferior izquierdo al ángulo superior derecho (Fig. 20).

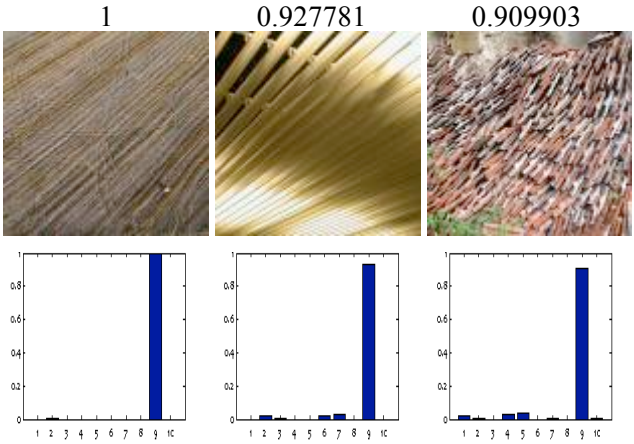


Figura 20. Imágenes tipificadas en el Aspecto-LE9 con mayor probabilidad y los histogramas de aspectos correspondientes.

Es interesante destacar el modo en que la probabilidad disminuye a la vez que la diagonal es menos perfecta y cómo aumentan en la tercera imagen el aspecto-LE4 Textura Heterogénea que se correspondería con las zonas superior e inferior de la imagen, y el aumento también de la probabilidad del aspecto-LE5 Vertical Irregular Texturada que pertenece a la parte derecha de la imagen en donde la diagonal ya es prácticamente una vertical.

10. Aspecto-LE10: Horizontal Vibrante. Este aspecto-LE10 Horizontal Vibrante presenta una textura de marcada horizontalidad de aspecto más bien rugoso y vibrante (Fig. 21).

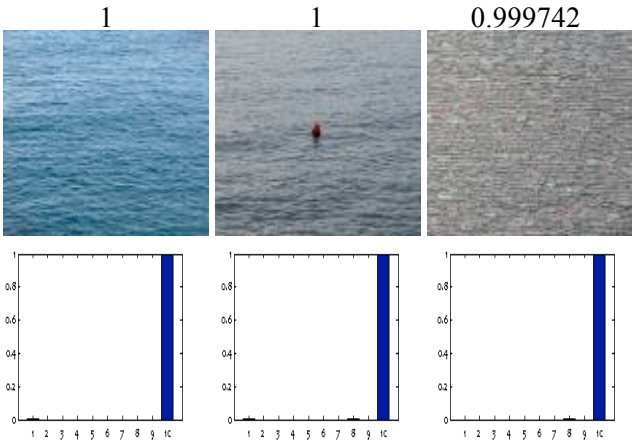


Figura 21. Imágenes tipificadas en el Aspecto-LE10 con mayor probabilidad y los histogramas correspondientes.

Aspectos de imágenes más entrópicas

La complejidad de esta categoría de imágenes es mucho mayor que la de imágenes poco entrópicas y el análisis de los resultados es complicado dado que ya no predomina de forma muy marcada un solo aspecto, sino que las imágenes están constituidas por multitud de ellos. A partir de ahora los aspectos pertenecientes a esta categoría de imágenes más entrópicas los nombraremos abreviados como HE (High entropy).

1. Aspecto-HE1: Figura-Fondo. Denominamos aspecto-HE1 Figura-Fondo al conjunto de imágenes en las que predomina una figura central en primer plano sobre un fondo bastante homogéneo. A pesar de tener probabilidades moderadamente superiores a 0,6, es un aspecto difícil de interpretar (Fig. 22).

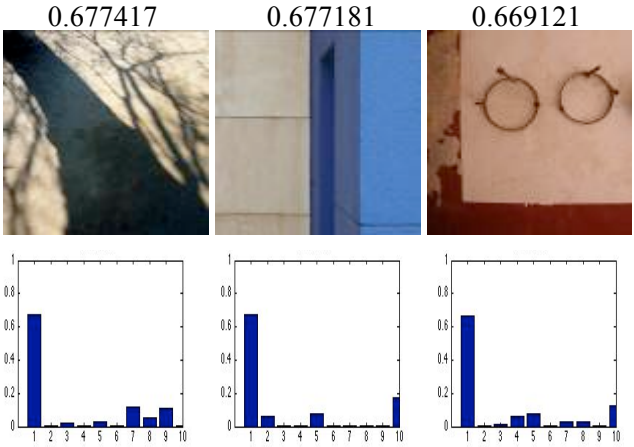


Figura 22. Imágenes tipificadas en el Aspecto-HE1 con mayor probabilidad y los histogramas de aspectos correspondientes.

2. Aspecto-HE2: Textura Homogénea Bipolar. Este aspecto pasado a escala de grises presenta dos tonalidades marcadas que dividen la imagen en dos secciones o polos tonales y/o estructurales de composición (Fig. 23).

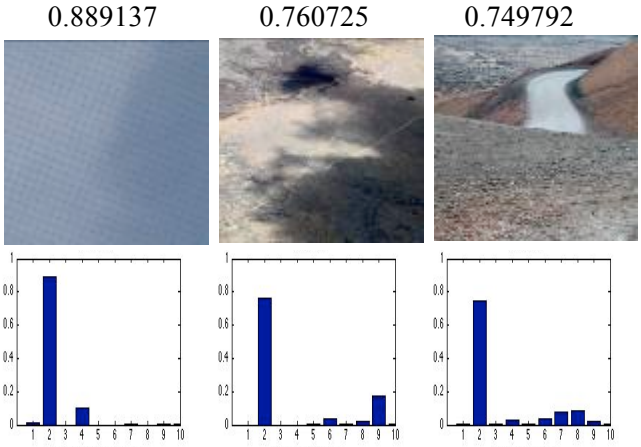


Figura 23. Imágenes tipificadas en el Aspecto-HE2 con mayor probabilidad y los histogramas de aspectos correspondientes.

A destacar en la primera imagen la componente del aspecto-HE4 Imagen Bipartita, en la segunda imagen el aspecto-HE9 Estructura Heterogénea Contrastada y en la tercera imagen la componente del aspecto-HE8 Planos Interseccionados.

3. Aspecto-HE3: Estructuras Verticales. En el aspecto-HE3 Estructuras Verticales se notan claramente unas franjas o columnas verticales en algún caso con una pequeña inclinación. Especialmente en las primeras imágenes se percibe un cierto predominio central en la composición (Fig. 24).

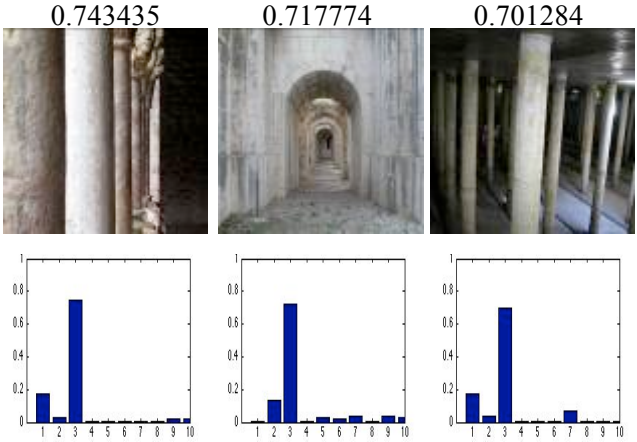


Figura 24. Imágenes tipificadas en el Aspecto-HE3 con mayor probabilidad y los histogramas de aspectos correspondientes.

En la primera y tercera foto aparece una componente importante del aspecto-HE1 Figura-Fondo posiblemente debido a las columnas que aparecen en el primer término.

4. Aspecto-HE4: Imagen Bipartita. El aspecto-HE4 Imagen Bipartita presenta claramente dos tipologías de textura dentro de la misma composición que suele coincidir con un paisaje de horizonte marcado (Fig. 25).

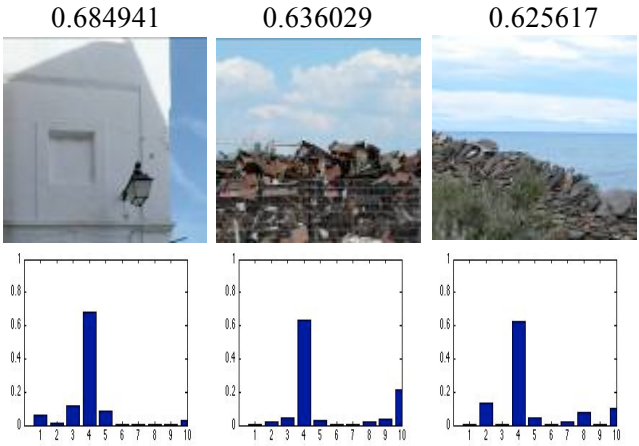


Fig. 25. Imágenes tipificadas en el Aspecto-HE4 con mayor probabilidad y los histogramas de aspectos correspondientes.

Nótese en la primera imagen las componentes de aspecto-HE3 Estructuras Verticales y aspecto-HE5 Cuadrícula.

5. Aspecto-HE5: Cuadrícula. Aspecto claramente perceptible en el que se puede visualizar perfectamente la estructura cuadricular de la composición (Fig. 26).

Cabe destacar cómo a medida que disminuye la probabilidad del aspecto-HE5 Cuadrícula, los cuadraditos son menos perfectos y a la vez aumenta la probabilidad del aspecto-HE7 Agrupaciones Rocosas.

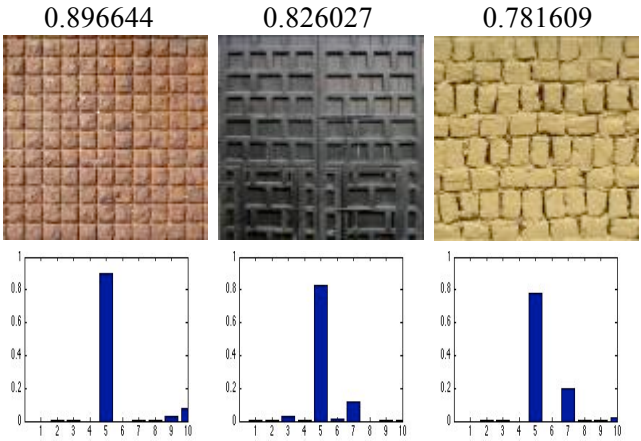


Figura 26. Imágenes tipificadas en el Aspecto-HE5 con mayor probabilidad y los histogramas de aspectos correspondientes.

6. Aspecto-HE6: Estructuras Horizontales. Este aspecto agrupa imágenes con una marcada tendencia a la horizontalidad en las estructuras predominantes. A diferencia del aspecto-LE3 Horizontal Amplia definido en las imágenes de baja entropía, éste no presenta únicamente líneas, sino más bien estructuras en forma de bandas (Fig. 27).

Es destacable la presencia en las 3 imágenes del aspecto-HE2 Textura Homogénea Bipolar (las 3 imágenes están compuestas básicamente por 2 texturas distintas) y la aparición en la segunda y tercera imágenes del aspecto-HE3 Estructuras Verticales debido a la componente vertical de los ladrillos.

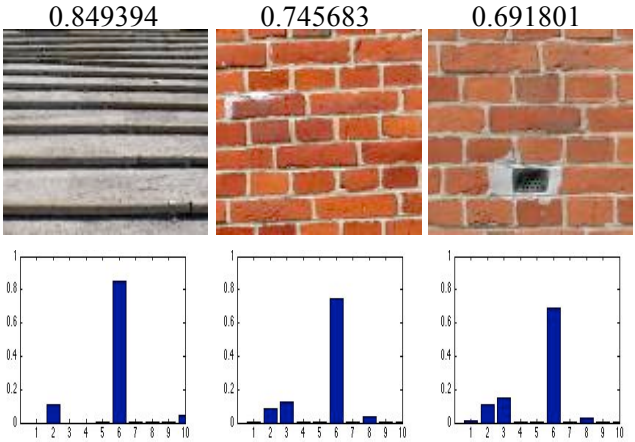


Figura 27. Imágenes tipificadas en el Aspecto-HE6 con mayor probabilidad y los histogramas de aspectos correspondientes.

7. Aspectos-HE7: Agrupaciones Rocosas. Bajo este aspecto percibimos como elemento común, en la mayoría de las imágenes, la piedra. Algunas presentan guijarros y otras restos de muros. Muchas de ellas están dispuestas por la mano del hombre en forma de muro o presentan un cierto orden natural que las caracteriza (Fig. 28).

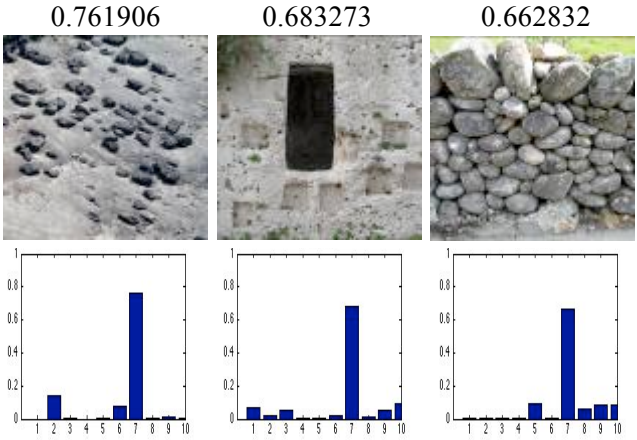


Figura 28. Imágenes tipificadas en el aspecto-HE7 con mayor probabilidad y los histogramas de aspectos correspondientes.

En la segunda imagen aparece el aspecto-HE1 Figura-Fondo, causado por la ventana en el primer término. Es interesante comparar la distribución de los aspectos HE5 Cuadrícula y HE7 Agrupaciones Rocosas de la tercera imagen de este aspecto con respecto a la tercera imagen del aspecto HE5 Cuadrícula. Tratándose en las dos imágenes de una agrupación de piedras, el sistema detecta la sutileza del distinto tipo de ordenamiento, más cuadriculado en el primer caso (Fig. 29).

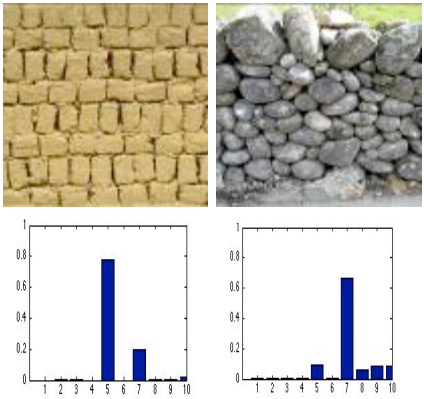


Figura 29. Primera imagen tipificada en el aspecto-HE5 y segunda imagen tipificada en el aspecto-HE7 y sus histogramas de aspectos correspondientes.

8. Aspectos-HE8: Líneas y Planos Inter-seccionados. En el aspecto-HE Líneas y Planos Inter-seccionados se recogen un conjunto de imágenes en las que predominan diferentes direccionalidades superpuestas en un mismo plano, que generalmente llenan todo el espacio (Fig. 30).

Destacable la componente de aspecto-HE6 Estructuras Horizontales de la imagen tercera, que puede explicarse si nos fijamos en las zonas sombreadas que crean unas bandas con tendencia horizontal.

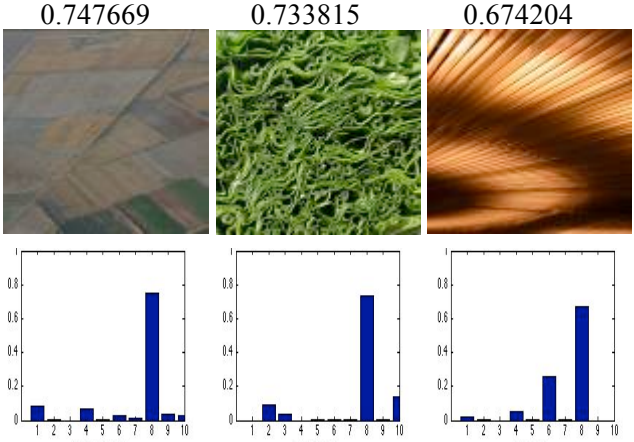


Figura 30. Imágenes tipificadas en el aspecto-HE8 con mayor probabilidad y los histogramas de aspectos correspondientes.

9. Aspecto-HE9: Estructura Heterogénea Contrastada. Este aspecto es muy compacto y aglutina un conjunto de imágenes en las que predominan estructuras volumétricas conseguidas por un claroscuro muy contrastado (Fig. 31).

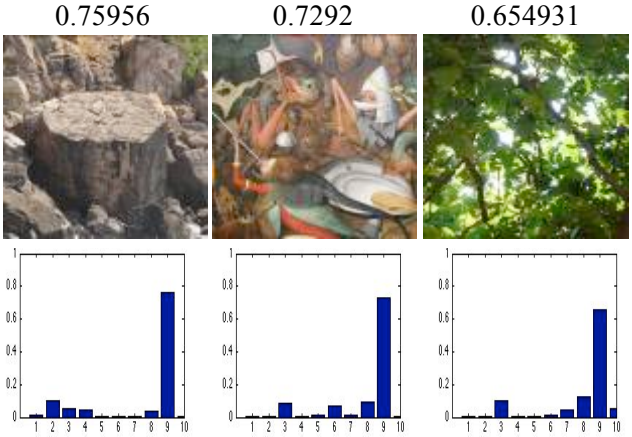


Figura 31. Imágenes tipificadas en el aspecto-HE9 con mayor probabilidad y los histogramas de aspectos correspondientes.

10. Aspecto-HE10: Figura Con Fondo Texturado. Este conjunto de imágenes presenta recuerda mucho al aspecto-HE1 Figura-Fondo, con la variante de que el fondo no es liso, sino texturado (Fig. 32).

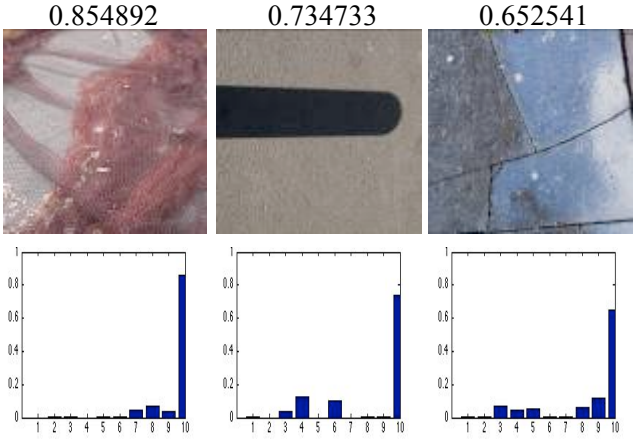


Figura 32. Imágenes tipificadas en el aspecto-HE10 con mayor probabilidad y los histogramas correspondientes.

Destacamos el componente de aspecto-HE4 Imagen Bipartita en la segunda imagen. Tampoco es un aspecto fácil de interpretar.

Conclusiones y comentarios

Es posible que algunos de los resultados expuestos anteriormente puedan parecer poco concluyentes para el lector no familiarizado con la visión artificial. Disponemos de la función HOG (histogram of oriented gradients) (Vedaldi & Fulkerson, 2008) que descompone una imagen en pequeñas celdas cuadradas y calcula un histograma de orientación en cada celda. La utilizamos a continuación en (Fig. 33) para generar una versión *pictórica* de las imágenes y así poder tener una aproximación visual de cómo *vería* el sistema computacional estas imágenes, en escala de grises y con vectores de gradiente, de cara a facilitar la comprensión de los resultados.

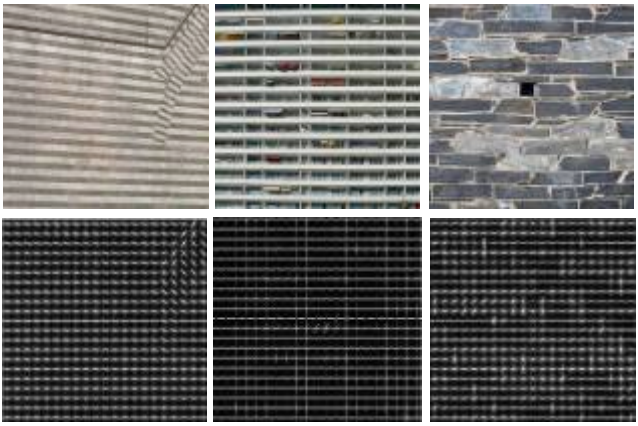


Figura 33. En la fila superior las 3 imágenes tipificadas en el aspecto-LE7 con mayor probabilidad y en la fila inferior las mismas imágenes procesadas con la función HOG.

En vista de las categorizaciones obtenidas podemos concluir que el sistema permitiría, dada una colección con gran cantidad de imágenes, realizar una importante preselección formal agrupándolas de forma más objetiva dado que la mirada artificial no está tan sujeta y condicionada como la percepción conceptual humana. El método de catalogación expuesto presentaría al autor nuevas relaciones, nuevos conjuntos difíciles de obviar, que podrían proporcionarle puntos de vista complementarios para futuros trabajos, en este sentido, el vocabulario personal del artista se vería enriquecido y más, conforme se vaya alimentando el sistema con nuevas imágenes. Todo esto será de gran ayuda en el proceso creativo, analítico, taxonómico y pedagógico de la obra.

Desde el punto de vista del espectador o del estudioso de la imagen, este tipo de sistemas puede ayudar a clarificar e interpretar, a hacer visibles matices de la imagen que de otra forma pasarían desapercibidos, y más en estos tiempos caracterizados por una saturación de imágenes, como nos apunta el artista visual Àlex Nogué (2013):

> "L'omnipresència de les imatges a la vida moderna ens ha fet desenvolupar uns mecanismes perceptius capaços d'actuar amb gran

> rapidesa. Hem aprés a captar les imatges immediatament i, emprant un sistema que fa intervenir la nostra potent memòria visual, en poques fraccions de segon som capaços de discernir-ne algun significat. Això només indica que el que s'ha vist encaixa amb el que ja havia estat vist i que encara persisteix en l'arxiu visual personal, però no vol dir que realment hàgim vist la imatge"[2]

No es la intención del presente trabajo abarcar la riqueza comunicacional de las obras del artista, tema muy complejo que estaría condicionado por múltiples factores como la psicología de la percepción estética y la relación que se establece entre la obra y el espectador, entre el que mira y el que es mirado (Hildebrand, 1988). Pero estamos convencidos de que estos sofisticados sistemas informáticos nos pueden ayudar a vislumbrar la cantidad y complejidad de procesos que lleva a cabo la mente del creador visual que persigue un objetivo estético, a la hora de decidir si una imagen pertenece o no a su vocabulario visual. Todo lo que pasa por su mente, la mayor parte procesos inconscientes, y le llevan a la decisión de capturar o no una instantánea.

Para concluir, citaremos a Joan Fontcuberta (2010), esperando haber contribuido modestamente a responder en parte a su pregunta:

> "Tanto desde la filosofía del arte como desde la semiótica se ha producido un esfuerzo para diagnosticar los rasgos que en una imagen permiten identificar al objeto representado. ¿Se trata de patrones basados en una mímesis gráfica objetiva y universal o por el contrario dependen de sistemas de representación culturales y objetivos? Una multiplicidad de hipótesis ha dado respuesta a estas cuestiones que en el fondo vienen impregnadas de una incertidumbre más profunda: la que atañe a nuestros modelos de construcción de la realidad… ¿Seria hoy posible diseccionar el concepto de semejanza según un criterio de lógica matemática?"[1]

Consideramos que la metodología de visión artificial tiene la potencialidad necesaria para detectar algunos de los trazos semánticos subyacentes en un gran conjunto de imágenes, incluso con significados muy abstractos como las mostradas en este trabajo, y sería capaz así de ayudar al artista a organizar su trabajo. Claro está que él siempre tendrá la última

palabra a la hora de decidir cuales de estos significados y sus combinaciones tienen el suficiente interés para materializarse en una obra de arte.

También en el ámbito académico este sistema puede ser de gran utilidad como recurso pedagógico para la mejora de la enseñanza artística. El docente y el discente pueden disponer de una herramienta de trabajo que les permita establecer unos criterios de análisis formales, planteados de manera objetiva, en un conjunto de imágenes o de obras. Este recurso aumentaría la posibilidad de conectar y de establecer relaciones entre imágenes de distintos estilos, épocas, temáticas, etc., ya que se convierte en un mecanismo autónomo y no dependiente de los hechos habitualmente tratados de carácter más historicista.

Agradecimientos

Mostramos nuestro profundo agradecimiento al Dr. M. Planas ya que sin su colaboración este trabajo no podría haberse realizado, al facilitarnos el acceso a su colección de imágenes y mostrarse en todo momento abierto y generoso en sus comentarios sobre los resultados, aportando valiosas sugerencias.

También agradecemos al Vicerrectorado de Política Docente y Científica de la Universidad de Barcelona 2011 – 2012, el soporte que nos ha prestado con la *Ayuda para la realización de proyectos de investigación pre competitivos en Ciencias Sociales y Humanas (APPCSHUM)*.

Notas

[1] Fontcuberta, J. (2010). p. 83.

[2] Nogué, A. ((2013). p.158.

Referencias

Arnheim, R. (1983). *Arte y percepción visual.* Madrid: Alianza forma.

Bosch, A., Zisserman, A. & Muñoz, X. (2006). Scene classification via PLSA. *In Proceedings of the European Conference on Computer Vision.* Graz, Austria.

Cover, T.M. & Thomas, J.A. (2006). *Elements of Information Theory* (2on ed.). New Jersey: John Wiley & Sons.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc. B*, 39, 1–38.

Enebral, J. (2009). *Detección y asociación automática de puntos característicos para diferentes aplicaciones*. Trabajo Final de Carrera. Escola Politècnica Superior de Castelldefels. Universitat Politècnica de Catalunya.

Fei-Fei, L. & Perona, P. (2005). *A Bayesian hierarchical model for learning natural scene categories. In Proc. CVPR.* San Diego, CA, USA.

Fontcuberta, J. (2010). *La cámara de pandora. La fotografí@ después de la fotografía.* Barcelona: Gustavo Gili.

Grauman, K. & Darrel, T. (2005). *The pyramid match kernel: Discriminative classification with sets of image features. In Proceedings of IEEE International Conference on Computer Vision (ICCV).* Beijing.

Hildebrand, A. von (1988). *El problema de la forma en la obra de arte.* Madrid: Visor.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis.. *Machine Learning*, 42,177-196.

Joachims, T. (1998). *Text categorization with support vector machines: Learning with many relevant features. In Proceedings of the European Conference on Machine Learning*. Springer-Verlag.

Lazebnik, S., Schmid, C. & Ponce, J. (2006). *Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. Computer Vision and Pattern Recognition, IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2, 2169-2178. doi.ieeecomputersociety.org/10.1109/CVPR.2006.68.

Lowe, D. G. (2004). Distinctive Image Features from Scale Invariant Keypoints. *Int. Journal of Computer Vision*, 60, 2, 91-110.

Nogué, A. (2013). *Dibuixar un arbre / Drawing a tree.* Barcelona: Comanegra.

Planas, M. A. (2014). *Miquel Planas.* Recuperado de http://www.miquelplanas.eu

Quelhas, P., Monay, F., Odobez, J.-M., Gatica-Perez, D., Tuytelaars & Van Gool, L. (2005). *Modeling scenes with local descriptors and latent aspects. Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05),* 1, 883-890. doi: 10.1109/ICCV.2005.152.

Vedaldi, A & Fulkerson, B. (2008). VLFeat - An open and portable library of computer vision algorithms. Retrieved from http://www.vlfeat.org

Willamowski, J., Arregui, D., Csurka, G., Dance, C., & Fan, L. (2004). Categorizing nine visual classes using local appearance descriptors. *In Proceedings of LAVS Workshop, in ICPR'04*, Cambridge.

Pilar Rosado: Estudiante de doctorado y becaria de Investigación del Departamento de Escultura de la Universidad de Barcelona.

Contact Address: Pau Gargallo 4, 08028 Barcelona (España).
E-mail address: pilarrosado@ub.edu

Eva Figueras: Profesora Titular del Departamento de Pintura de la Universidad de Barcelona.

Contact Address: Pau Gargallo 4, 08028 Barcelona (España).
de la E-mail address: efigueras@ub.edu

Ferran Reverter: Profesor Lector del Departamento de Estadística de la Universidad de Barcelona e Investigador Postdoc del Centre de Regulació Genòmica del Parc de Recerca Biomèdica de Barcelona.

Contact Address: Avda. Diagonal 643, 08028 Barcelona (España).
E-mail address: freverter@ub.edu

Anexos

Anexo A

Obtención de descriptores locales

a. Extracción de características locales (SIFT Descriptors)

El descriptor *Scale Invariant Feature Transform* (SIFT) fue desarrollado por (Lowe, 2004) como un algoritmo capaz de detectar puntos característicos (*keypoints*) estables en una imagen. Estos puntos son invariantes frente a diferentes transformaciones como traslación, escala, rotación, iluminación y transformaciones afines. Originalmente fue desarrollado para el reconocimiento de objetos en general y para realizar la alineación de imágenes. El algoritmo SIFT se compone principalmente de cuatro etapas que se describen siguiendo la implementación de (Lowe, 2004):

a. Detección de extremos en el Espacio de Escala: La primera etapa del algoritmo realiza una búsqueda sobre las diferentes escalas y dimensiones de la imagen identificando los candidatos a *keypoints*. Esto se lleva a cabo mediante la función DoG (Difference-of-Gaussian).
b. Localización de los *keypoints*: Se seleccionan los *keypoints* a partir del conjunto de candidatos encontrados, aplicando una medida de estabilidad sobre todos ellos para descartar los que no sean adecuados.
c. Asignación de la orientación: Se asignan una o más orientaciones a cada *keypoint* basándose en las direcciones locales presentes en la imagen gradiente. Todas las operaciones posteriores serán realizadas sobre los datos transformados según la orientación, escala y localización dentro de la imagen, lo que nos proporcionará la invariancia parcial a distorsiones de forma así como a cambios de iluminación.
d. Descriptor del *keypoint*: La última etapa hace referencia a la representación de los *keypoints* como una medida de los gradientes locales de la imagen en las proximidades de dichos puntos clave y

respecto de una determinada escala. Cada punto de interés corresponde a un vector de características compuesto por 128 elementos.

A continuación detallaremos las etapas anteriores (Enebral, 2009)

b. Detección de extremos en el Espacio de Escala

La primera fase del algoritmo es la encargada de buscar un primer conjunto de *keypoints* de la imagen candidatos a poder ser identificados de forma repetida bajo diferentes vistas del mismo objeto. La detección de ubicaciones que son invariantes frente cambios de escala de la imagen se puede lograr mediante la búsqueda de características estables en todas las escalas posibles, utilizando una función continua de escala conocida como función Espacio de Escala.

La función Espacio de Escala de una imagen se define como la función, $L(x, y, \sigma)$, que resulta de la convolución de una función Gausiana, $G(x, y, \sigma)$, con la imagen original $I(x, y)$:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y)$$

donde $*$ denota el operador convolución en x y y, y

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} \exp(-\frac{x^2 + y^2}{2\sigma^2})$$

El cálculo de todo el espacio $L(x, y, \sigma)$ se lleva a cabo construyendo una pirámide gausiana (Fig. 34), convolucionando con diferentes filtros $G(x, y, \sigma)$ al variar el parámetro σ. Las imágenes de la pirámide gausiana se distribuyen según los términos siguientes:

- Octava: Imágenes del espacio $L(x, y, \sigma)$ de igual tamaño que difieren en el parámetro de filtrado σ con el que han sido obtenidas.
- Escala: Imágenes del espacio $L(x, y, \sigma)$ filtradas con el mismo parámetro σ pero con diferentes tamaños.

Para mejorar la estabilidad de los puntos de interés que se obtendrán más adelante, es conveniente realizar un pre-procesado. La imagen original $I(x, y)$ se suaviza mediante un filtrado gausiano con $\sigma_0 = 0.5$ y

posteriormente se re-escala con un factor 2 usando interpolación lineal. La imagen resultante, al doblar su tamaño, le corresponderá un valor $\sigma_1 = 1$ y es ésta la que se utilizará como imagen inicial para construir la pirámide.

Los diferentes valores del parámetro σ con los que se configura la pirámide tienen que verificar en cada octava la siguiente condición: el penúltimo, σ_4 en este caso, ha de ser el doble que el primero, $\sigma_1 = 1$. Por consiguiente, dividiremos cada octava en intervalos múltiplos de k

$$k = 2^{\frac{1}{(\text{num.escalas})-2}} = 2^{\frac{1}{3}}$$

entonces

$$\sigma_i = k^{i-1} = 2^{\frac{i-1}{3}} \quad i = 1,...,5$$

Una vez terminada la primera octava, se elige la imagen con $\sigma_4 = 2$ como imagen inicial de la siguiente octava, de esta manera, al re-escalarla a la mitad su factor de filtrado vuelve a ser $\sigma_1 = 1$. Este proceso se va repitiendo hasta completar toda la pirámide (Fig. 34).

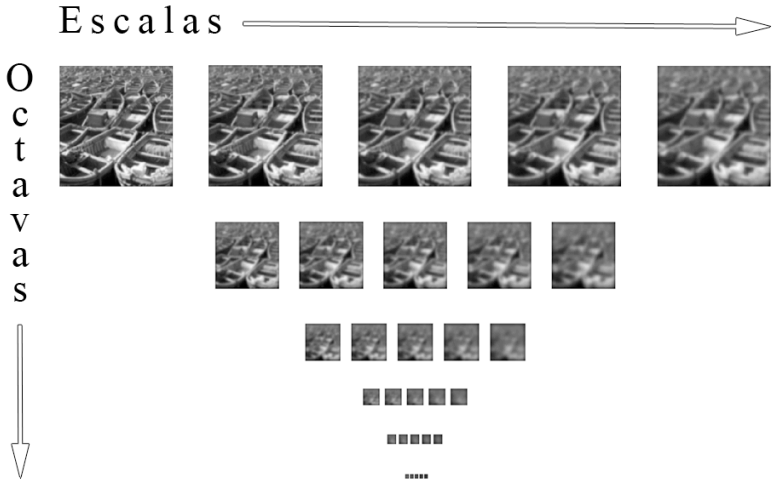


Figura 34. Pirámide gausiana.

c. Localización de los keypoints

Para detectar puntos de interés estables en el Espacio de Escala utilizamos la función DoG (Difference-of-Gaussian), definida por:

$$D(x,y,\sigma) = \big(G(x,y,k\sigma) - G(x,y,\sigma)\big) * I(x,y)$$

$$= L(x,y,k\sigma) - L(x,y,\sigma)$$

Obsérvese que la función DoG resulta de la función $L(x,y,\sigma)$ calculada en la etapa anterior. La obtención de la función DoG no comporta un incremento considerable del coste computacional total, ya que se calcula simplemente restando imágenes vecinas de una misma octava. En la pirámide DoG tendremos cuatro imágenes-resta por octava (Fig. 35).

Figura 35. Imágenes-resta de una primera octava. La pirámide DoG se completa obteniendo las imágenes-resta de las sucesivas octavas.

A partir de los cálculos anteriores, se hallarán los máximos y mínimos locales del espacio $D(x,y,\sigma)$. En esta etapa cada uno de los píxeles de cada imagen de la pirámide se compararán con sus ocho vecinos de la propia imagen y con los nueve vecinos anteriores y posteriores de escala (Fig. 36).

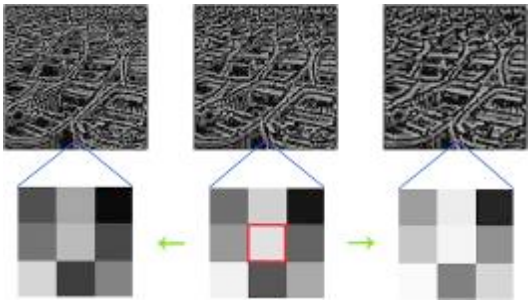


Figura 36. Vecinos anterior y posterior de escala

Un punto quedará seleccionado como *keypoint* sólo si es mayor que sus 26 vecinos o menor que todos ellos. Observamos que sólo se podrán detectar *keypoints* en escalas centrales de $D(x, y, \sigma)$ pues no existen imágenes vecinas en las escalas laterales.

En esta fase del método SIFT se centra en almacenar toda la información disponible de cada *keypoint*. Es decir, para cada punto de interés encontrado se guardará a qué escala y octava de la pirámide pertenece, y su posición, es decir la fila y la columna, dentro de la imagen correspondiente.

d. Asignación de la orientación

En esta etapa calcularemos las orientaciones de cada punto de interés. Una vez las tengamos podremos construir descriptores invariantes a la rotación, ya que éstos serán referenciados a sus respectivas orientaciones.

Alrededor del punto donde vamos a determinar la orientación definimos una región de 16x16 píxeles y a cada uno de los píxeles se le calcula su gradiente (Fig. 37a y Fig. 37b). El gradiente viene determinado por su módulo $m(x, y)$ e inclinación $\theta(x, y)$, ambos se calculan utilizando deferencias entre pixeles:

$$m(x,y) = \sqrt{\big(L(x+1,y) - L(x-1,y)\big)^2 + \big(L(x,y+1) - L(x,y-1)\big)^2}$$

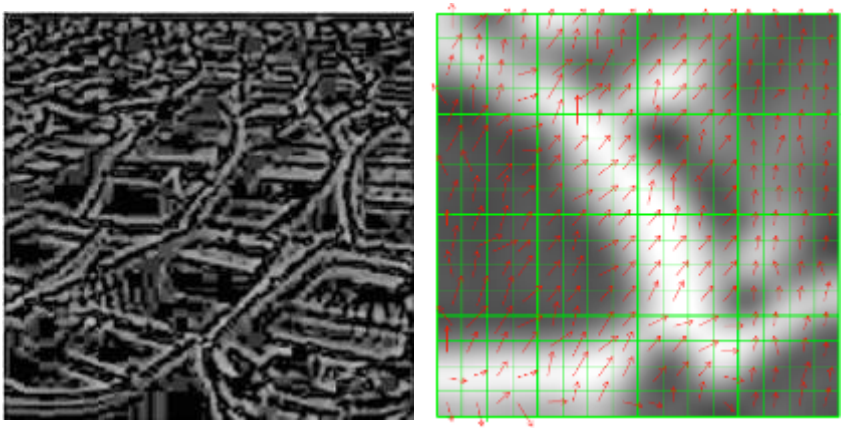
$$\theta(x,y) = \tan^{-1}\left(\frac{L(x,y+1) - L(x,y-1}{L(x+1,y) - L(x-1,y)}\right)$$

La imagen empleada para obtener $m(x,y)$ y $\theta(x,y)$ será la imagen de la pirámide $L(x,y,\sigma)$ (Fig. 34) donde se detectó el punto de interés que está siendo analizado.

Después de realizar el proceso anterior, se agrupará la información en forma de histograma, uno para cada punto de interés. De esta manera cada histograma de orientaciones estará formado por 36 bins o clases para completar el rango total de 360º (Fig. 37c). A medida que se añade al histograma cada orientación $\theta(x,y)$, dicho valor se pondera por su módulo $m(x,y)$ y por una ventana circular gausiana con valor σ igual a 1.5 veces la escala del punto de interés. Los motivos principales para realizar estas dos ponderaciones son: dar mayor peso a las orientaciones con módulos elevados y mayor importancia a los puntos cercanos al punto de interés.

El bin modal de cada histograma corresponderá la dirección dominante de los gradientes locales, y por lo tanto la orientación final del punto de interés.

a) b)



c)

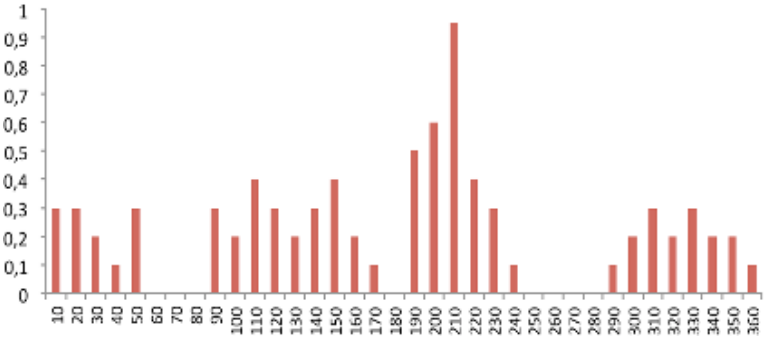


Figura 37. a) *Keypoint*. b) Región de 16x16 píxeles alrededor del *keypoint* y gradiente. c) Histograma de orientaciones. La orientación del *keypoint* corresponde al valor máximo.

e. Descriptores de los keypoints

Las etapas anteriores han dotado a los puntos de interés seleccionados de invariancia respecto de la orientación, escala y localización respecto de la imagen. En esta última etapa se crea un vector de características para cada uno de los puntos de interés que contiene una estadística local de las orientaciones del gradiente.

El proceso parte de las regiones 16x16 del apartado anterior ya multiplicadas por la ventana gausiana con σ igual a 1.5 veces la escala del punto de interés (Fig. 38a). Cada una de estas regiones se divide en subregiones de 4x4 píxeles con el objetivo de resumir toda esa información en pequeños histogramas de sólo 8 bins, es decir, 8 orientaciones. Previamente a la realización de esta reconfiguración, cada gradiente de la ventana 16x16 se rota tantos grados como especifique la orientación del punto de interés (calculada en la etapa anterior, Fig. 38c), y así será independiente a la inclinación de la imagen.

Para cada punto de interés ahora pasaremos a tener 16 pequeños histogramas de 8 bins cada uno de ellos. Para evitar cambios abruptos entre las fronteras de las subregiones, cada una es filtrada de nuevo por una ventana circular gausiana (en esta ocasión de tamaño 4x4) con un factor σ = 0.5 x escala del punto de interés (Fig. 38b).

Cada uno de los histogramas se compone de 8 bins, que almacenan las orientaciones posibles proporcionales a 45 grados donde la magnitud de cada flecha representa el valor acumulado para cada bin. Por lo tanto se obtienen 16 histogramas respecto de las orientaciones de los puntos de cada región para cada uno de los puntos de interés.

Finalmente el descriptor de cada punto de interés está formado por un vector que contiene los valores de las 8 orientaciones de los 4x4 histogramas componiendo un vector de características de 4 x 4 x 8 = 128 elementos (Fig. 38c).

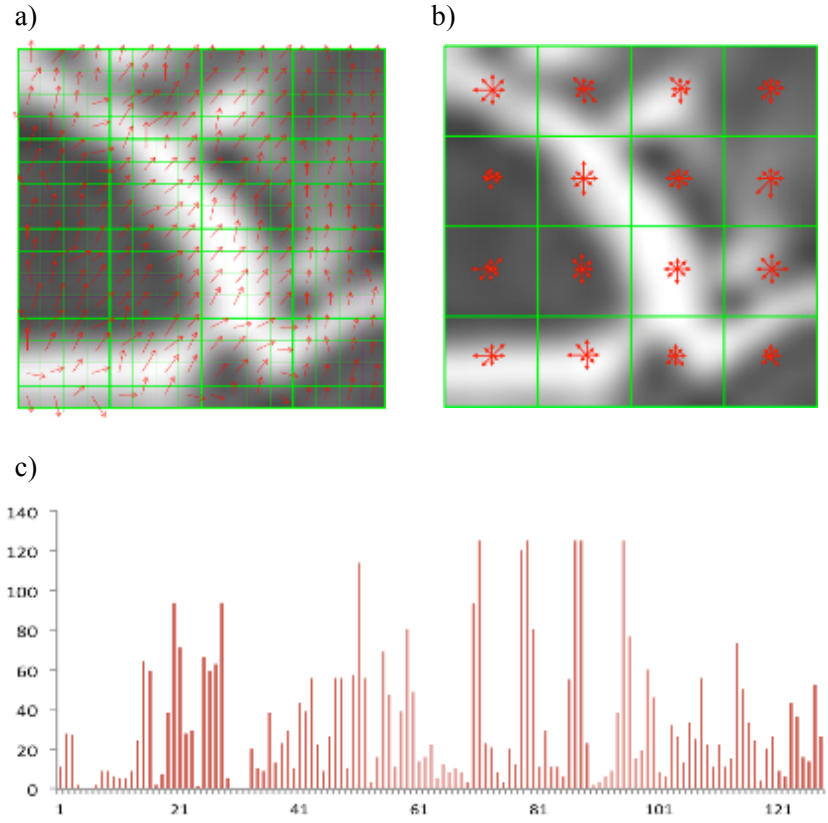


Figura 38. a) Región de 16x16 píxeles alrededor del *keypoint* y gradiente. b) Subregiones de 4x4 píxeles con histogramas de sólo 8 orientaciones c) Descriptor.

Las etapas anteriores aseguran que los descriptores obtenidos sean invariantes frente a la luminosidad. Esto es consecuencia de que los gradientes están calculados mediante diferencias entre píxeles vecinos. De esta manera sumar una constante de luz a la imagen no influirá en el resultado final.

Para lograr invariancia frente a cambios de contraste hay que normalizar a la unidad cada uno de los 'sub-histogramas' del total de 16 que tiene cada descriptor.

Por la saturación de la cámara o por cambios de iluminación sobre superficies puede producirse una variación no lineal de luz.
Ante esta variación para reducir sus efectos impondremos un umbral superior de 0.2 a los histogramas normalizados y posteriormente se renormalizará de nuevo a la unidad. Efectuadas estas modificaciones, el proceso de construcción de los descriptores queda completado.

En algunos estudios (Lazebnik, Schmid & Ponce, 2006; Fei-Fei & Perona, 2005) el cálculo de los descriptores locales SIFT, en vez de realizarse en los puntos de interés, se efectúa en los nodos de una malla regular sobreimpuesta en la imagen (Fig. 39). Este enfoque es preferible con el fin de mejorar la capacidad de discriminación en implementaciones orientadas a la clasificación de escenas.

Figura 39. Imagen con una malla de 10 x 10. Los nodos contienen los valores de las 8 orientaciones de los histogramas 4x4. Para cada nodo resulta un vector de características, con 4 x 4 x 8 = 128 elementos

Anexo B

Vocabulario visual

a. Construcción del vocabulario visual

El punto de partida para la construcción de un vocabulario visual es el conjunto de descriptores $F = \{f_i : i = 1,...,N_F\}$ de la colección de imágenes $D = \{d_1,...,d_N\}$ y el punto al cual queremos llegar es un vocabulario de "visual terms" $V = \{v_1,...,v_M\}$. Utilizaremos la expresión: palabra visual como equivalente a la expresión inglesa "visual term". Cada imagen d_i ha quedado descrita mediante los descriptores SIFT, denotemos genéricamente por f un descriptor. Considerando toda la colección de imágenes tenemos por tanto una gran colección de descriptores. La construcción del vocabulario requiere la cuantización de cada descriptor local f en su respectiva palabra visual v_i de acuerdo con la siguiente regla de asignación:

$$f \to Q(f) = v_i \Leftrightarrow \text{dist}(f, v_i) \le \text{dist}(f, v_j), \quad j = 1,...,M \qquad (1)$$

donde $dist(\cdot,\cdot)$ es una función distancia.

Si indicamos con S el espacio de los descriptores, en nuestro caso al tener descriptores 128 dimensionales podemos asumir que $S \equiv \mathrm{R}^{128}$. Una vez fijado el vocabulario $V = \{v_1,...,v_M\}$, S queda dividido en M regiones $S = \{S_1,...,S_M\}$ de acuerdo con:

$$S_i = \{f \in S : Q(f) = v_i\}$$

La construcción del vocabulario se realiza mediante agrupación (clustering). Más específicamente, aplicamos el algoritmo k-means a un conjunto representativo de descriptores locales extraídos de la colección de imágenes y tomaremos como palabras visuales los vectores de medias de cada clúster. Usamos la distancia euclidiana en los procesos de agrupación

y cuantización y elegimos el número de clústeres dependiendo del tamaño deseado de vocabulario.

El algoritmo k-means establece una partición o agrupación del conjunto de descriptores

$$F = \{f_i : i = 1,..., N_F\}$$

En M subconjuntos disjuntos S_i que contienen los descriptores que minimizan la función de error cuadrático:

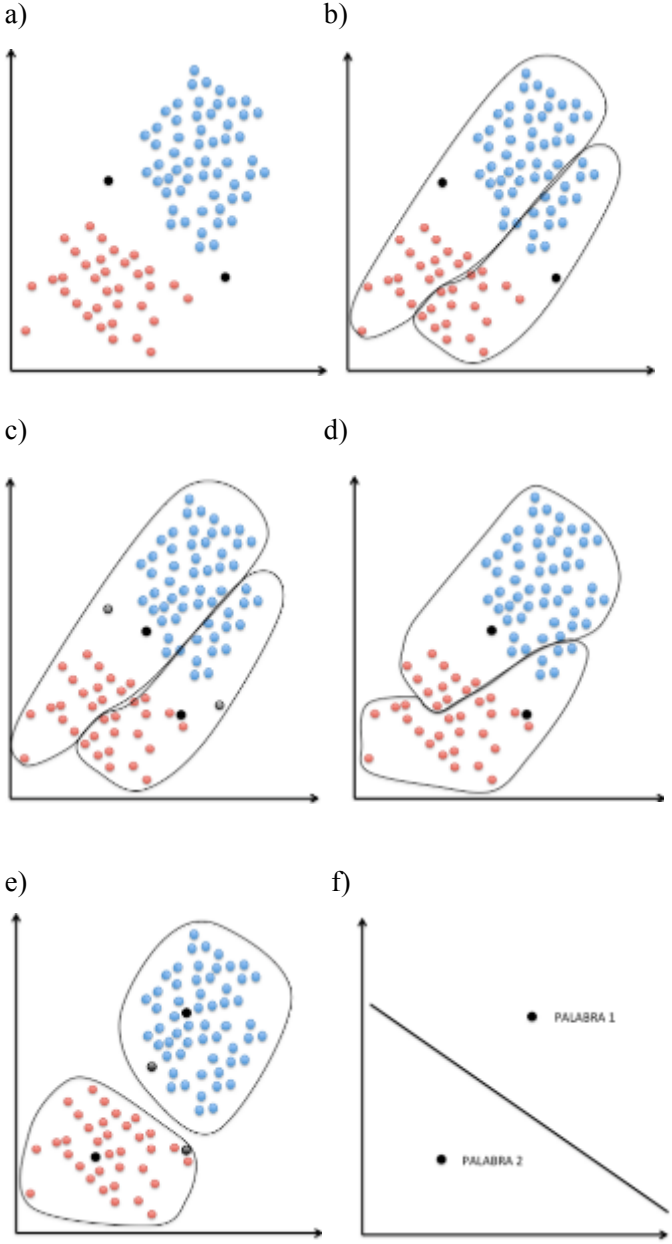
$$J(F) = \sum_{i=1}^{M} \sum_{j \in S_i} (f_j - \mu_i)^2$$

Donde μ_i denota el vector de medias (centroide) del subconjunto de descriptores S_i.

El algoritmo busca la partición mediante la iteración de dos etapas. La primera etapa consiste en asignar cada descriptor al centroide más cercano. En la segunda atapa se recalculan los centroides de cada región, calculando el vector de medias de los descriptores que han sido asignados a cada región (Fig. 40).

b. Algoritmo K-means

- Establecer al azar M centroides iniciales.
- Asignar cada descriptor al subconjunto S_i que tenga el centroide μ_i más cercano, de acuerdo con la fórmula 1.
- Recalcular el valor del centroide μ_i mediante el vector de medias de los descriptores asignados a S_i.
- Repetir los pasos 2 y 3 hasta que los valores de los centroides μ_i no se modifiquen.


a)
b)
c)
d)
e)
f)
PALABRA 1
PALABRA 2

g)

Figura 40. A modo de ejemplo, supongamos el caso de descriptores bidimensionales y de dos palabras visuales. El algoritmo k-means establecerá una partición del espacio en dos regiones cada una asociada a una palabra. a) Supongamos que los descriptores de la colección de imágenes configuran dos grupos separados (azul y rojo). El algoritmo empieza estableciendo dos centroides al azar (negro), b) Asignamos cada descriptor al centroide más cercano. c) Recalculamos los nuevos centroides de los grupos formados en la etapa anterior. d) Repetimos la asignación de los descriptores al centroide más cercano. e) El procedimiento prosigue recalculando los nuevos centroides. f) El proceso iterativo se detiene cuando no se produce cambio en los centroides. g). Ilustra la partición del espacio de descriptores en el caso de una vocabulario de más palabras. Dado un descriptor f calcularemos el centroide más cercano, y le corresponderá la palabra representada por dicho centroide.

Dada una imagen d con un conjunto de descriptores

$$F(d) = \{f_j : j = 1,..., N_{F(d)}\}$$

podemos usar los centroides obtenidos en el algoritmo k-means para atribuir la palabra visual v_i a todo descriptor f_j para el que el centroide más cercano sea μ_i. Una vez completada la atribución obtenemos la representación BOV de la imagen:

$$h(d) = (h_1(d),..., h_M(d)) \quad , \quad h_i(d) = n(d, v_i)$$

donde $n(d, v_i)$ indica la frecuencia de la palabra visual v_i en la imagen d .

Anexo C

Representación de aspectos latentes

Vamos a explicar el modelo en términos de imágenes, palabras visuales y aspectos. Disponemos de una colección de imágenes $D = \{d_1,...,d_N\}$ y un vocabulario de palabras visuales $V = \{v_1,...,v_M\}$. Podemos resumir las observaciones en una tabla $N \times M$ de frecuencias $n(d_i, v_j)$, donde $n(d_i, v_j)$ indica la frecuencia con que la palabra visual v_j ocurre en la imagen d_i. PLSA es un modelo estadístico generativo que asocia una variable latente $z_l \in \{z_1,...,z_K\}$ con cada observación, entendiendo por observación la ocurrencia de una palabra visual en una imagen dada. Estas variables, normalmente llamadas aspectos, se utilizan para construir un modelo de probabilidad conjunta sobre las imágenes y las palabras visuales, definido por:

$$P(d_i, v_j) = P(d_i) \sum_{k=1}^{K} P(v_j|z_k) P(z_k|d_i)$$

donde $P(d_i)$ indica la probabilidad de d_i $P(v_j|z_k)$ indica la probabilidad condicionada de una palabra visual específica condicionada al aspecto latente z_k, y $P(z_k|d_i)$ indica la probabilidad condicional específica de cada imagen. El PLSA introduce un principio de independencia condicional: asume que la ocurrencia de una palabra visual v_j es independiente de la imagen d_i en la que esta, dado un aspecto z_k.

La estimación de las probabilidades del modelo PLSA se llevan a cabo mediante el máximo de la verosimilitud utilizando la colección de imágenes $D = \{d_1,...,d_N\}$. La optimización se resuelve mediante el algoritmo EM (Dempster, Laird, & Rubin, 1977). El algoritmo EM alterna dos etapas. En la etapa E se calculan las probabilidades a posteriori para los aspectos latentes basándonos en las estimaciones actuales de las probabilidades del modelo, en la etapa M las probabilidades del modelo se actualizan maximizando la llamada *"expected complete data log-likelihood"*:

Etapa E

$$P(z_k|d_i,v_j)=\frac{P(v_j|z_k)P(z_k|d_i)}{\sum_{l=1}^{K}P(v_j|z_l)P(z_l|d_i)}$$

Etapa M

$$P(v_j|z_k)=\frac{\sum_{i=1}^{N}n(d_i,v_j)P(z_k|d_i,v_j)}{\sum_{m=1}^{M}\sum_{i=1}^{N}n(d_i,v_m)P(z_k|d_i,v_m)}$$

$$P(z_k|d_i)=\frac{\sum_{j=1}^{M}n(d_i,v_j)P(z_k|d_i,v_j)}{n(d_i)}\quad,\quad n(d_i)=\sum_{j=1}^{M}n(d_i,v_j)$$

Las etapas E y M se alternan hasta que se alcanza una cierta condición de terminación. El proceso iterativo se inicia asignando valores aleatorios al conjunto de probabilidades $P(z_k|d_i)$ y $P(v_j|z_k)$.

Como consecuencia del proceso anterior obtenemos una nueva representación para las imágenes de la colección basada en la distribución de aspectos,

$$a(d_i)=\left(P(z_1|d_i),\ldots,P(z_K|d_i)\right)$$

De hecho, también es posible hallar la distribución de aspectos para una imagen cualquiera que no forme parte de la colección inicial (Quelhas, Monay, Odobez, Gatica-Perez, Tuytelaars & Van Gool, 2005; Bosch,

Zisserman & Muñoz, 2006). Basta recurrir de nuevo al algoritmo EM antes descrito pero en este caso en la etapa M sólo se actualizan las probabilidades $P(z_k|d)$ y las probabilidades $P(v_j|z_k)$, independientes de la imagen, estimadas a partir de la colección en la fase de aprendizaje, se mantienen fijas.

Si bien la representación de imágenes basada en aspectos se puede usar como punto de entrada para alimentar un clasificador de escenas, nosotros vamos a centrarnos en la utilización de dicha representación para la ordenación o ranking de imágenes basada en la distribución de aspectos subyacentes. Dado un aspecto z, las imágenes pueden ordenarse según los valores:

$$P(d|z) = \frac{P(z|d)P(d)}{P(z)} \propto P(z|d)$$

de esta manera, una vez estimados los valores de $P(z_k|d)$ $k = 1,\ldots,K$, para una imagen dada d, podemos ordenarlos y tener una medida objetiva de la asociación entre la imagen y cada uno de los aspectos. En consecuencia, asociaremos la imagen al aspecto con mayor probabilidad.