2nd Art, Science, City International Conference ASC2015
Valencia, Universitat Politècnica de València, 22-23 octubre 2015
DOI: http://dx.doi.org/10.4995/ASC/ASC15.2015.1985

LA IMAGEN TOMA LA PALABRA: CONSTRUCCIÓN DE UN VOCABULARIO VISUAL

Pilar Rosado Rodrigo
Eva Figueras Ferrer
Miguel Planas Roselló
Universidad de Barcelona, Facultad de Bellas Artes, Departamentos de Pintura y Escultura
prrforma@gmail.com, efigueras@ub.edu, maplanasrossello@ub.edu,
Ferran Reverter Comes
Universidad de Barcelona, Facultad de Biología, Departamento de Estadística, freverter@ub.edu

Resumen

Es secular la pugna entre escritura e imagen. En la actualidad existen numerosos indicios de que es necesario un retorno hacia la imagen: es importante encontrar analogías con el lenguaje que puedan aplicarse a la información visual. Si ha sido posible descomponer el lenguaje en elementos y estructuras ¿sería posible hacerlo también con las imágenes?

La imagen digital nos brinda la oportunidad de describir las formas en términos matemáticos y así pone a nuestro alcance la posibilidad de descifrar el problema del significado contenido en la imagen.

Nuestra investigación propone la búsqueda de analogías formales en grandes colecciones de imágenes de obras de artista abstractas, basada únicamente en su contenido visual y sin poyo de anotación textual alguna.

Se ha programado un algoritmo de descripción de imágenes utilizado en visión artificial cuyo enfoque consiste en colocar una malla regular de puntos de interés en la imagen y seleccionar alrededor de cada uno de sus nodos una región de píxeles para la que se calcula un descriptor que tiene en cuenta los gradientes de grises encontrados.

Los descriptores de toda la colección de imágenes se pueden agrupar en función de su similitud y cada grupo resultante pasará a determinar lo que llamamos. El total de "palabras visuales" de la colección de imágenes genera un vocabulario visual. El método se denomina Bag-of-Words (bolsa de palabras) porque representa una imagen como una colección desordenada de características visuales locales. Teniendo en cuenta la frecuencia con que cada "palabra visual" ocurre en cada imagen, aplicamos el pLSA (Probabilistic Latent Semantic Analysis), un modelo estadístico que clasificará de forma totalmente automática las imágenes según su categoría formal.

De esta manera se espera obtener una herramienta de utilidad tanto en la producción artística como en el análisis de obras de arte.

Palabras-clave: PALABRA VISUAL, VISIÓN ARTIFICIAL, MODELO BAG-OF- WORDS, CBIR (RECUPERACIÓN DE IMÁGENES POR CONTENIDO), PLSA (ANÁLISIS PROBABILÍSTICO DE ASPECTOS LATENTES), ARTE DIGITAL,

Abstract

Conflict between writing and image is ancient. At present, there are many evidences that a return to the image is needed. To find analogies between language and visual information is important. If it has been possible to decompose language in elements and structures, why not with images?

The opportunity offered by digital image to describe lines and shapes in mathematical terms provides us the ability to decipher the problem of meaning contained in image.

The objective of our research is to develop a series of computer vision programs to search for analogies in large datasets—in this case, collections of images of abstract paintings—based solely on their visual content without textual annotation.

We have programmed an algorithm based on a specific model of image description used in computer vision. This approach involves placing a regular grid over the image and selecting a pixel region around each node. Dense features computed over this regular grid with overlapping patches are used to represent the images. Analysing the distances between the whole set of image descriptors we are able to group them according to their similarity and each resulting group will determines what we call "visual words". Considering the whole collection of images, the total collection of "visual words" will define his "visual vocabulary". This model is called Bag-of-Words representation of an image because does not contain information concerning the spatial relationships among the visual words which make it up. Given the frequency with which each visual word occurs in each image, we apply the method pLSA (Probabilistic Latent Semantic Analysis), a statistical model that classifies fully automatically, without any textual annotation, images according to their formal patterns.

In this way, the researchers hope to develop a tool both for producing and analysing works of art.

Keywords: VISUAL WORD, ARTIFICIAL VISIÓN, BAG-OF-WORDS MODEL, CBIR (CONTENT-BASED IMAGE RETRIEVAL), PLSA (PROBABILISTIC LATENT SEMANTIC ANALYSIS), DIGITAL ART

1. INTRODUCCIÓN

La creación del relato hablado nos permitió representar conceptos mediante vocablos diferenciados. Posteriormente, con la invención del lenguaje escrito, desarrollamos diferentes formas de simbolizar nuestros pensamientos. Así, mediante ideas estructuradas, las bibliotecas aumentaron enormemente la capacidad de nuestros cerebros de retener y expandir nuestra base de conocimientos.

Flusser (2009) propone que en la cultura humana se han producido dos acontecimientos fundamentales; el primero la "invención de la escritura lineal" alrededor de la mitad del segundo milenio antes de Cristo, y el segundo la "invención de las imágenes técnicas", en el momento actual. Esta visión de la historia nos sitúa en la pugna entre escritura e imagen, entre dos "contenedores de significados" que codifican y contienen el tiempo de manera diferencial; este autor nos habla del tiempo circular de la magia en las imágenes y del tiempo lineal de la historia en los escritos.

La saturación de imágenes a la que nos vemos sometidos en la actualidad: la supremacía de internet, las redes sociales, el abaratamiento de las cámaras digitales y su implantación en los teléfonos móviles, etc., contribuyen a la ubiquidad de la imagen en nuestra realidad. La inmediatez y las prisas favorecen también que nuestra mirada se haya vuelto más superficial. Ha dejado de ser un problema el almacenamiento de la información, disponemos de cantidades ingentes de contenidos, incluso a nivel doméstico, en nuestras casas, guardados en los dispositivos electrónicos que tenemos a nuestro alcance. Actualmente el verdadero problema es cómo acceder a ellos, pues dependemos de la mirada parcial que nos proporcione el índice de acceso que utilicemos.

> "La forma tradicional de visualizar la información ya no es válida. Necesitamos técnicas que nos permitan observar los vastos universos de los media para poder detectar rápidamente multitud de patrones de interés. Estas técnicas tienen que ser compatibles con la capacidad de procesamiento de información del ser humano y, al mismo tiempo, conservar una cantidad suficiente de detalles de las imágenes originales, video, audio o experiencias interactivas para permitir su estudio" (Manovich 2012, 1).

La visión artificial proporciona alternativas en este sentido y por ello se invierten importantes esfuerzos en esta dirección. Es un hecho que existe una sintaxis visual, unas líneas generales de construcción de composiciones, elementos básicos y mensajes visuales que se pueden comprender y aprender, seas artista o no (Dondis, 1984). Captamos información visual de muchas formas y a ello le afecta tanto la fisiología perceptiva como nuestro propio movimiento o estado de ánimo. A pesar de que existen diferencias a nivel individual y colectivo, existe un sistema perceptivo visual que todos los seres humanos compartimos. La visión artificial aplicada al estudio de la imagen digital ofrece la posibilidad de detección de semejanzas formales susceptibles de ser utilizadas para la extracción de información visual.

En el presente estudio proponemos conseguir una clasificación automática de imágenes digitales de obras de artista abstractas basada únicamente en su contenido semántico, sin necesidad de anotación textual alguna, que sea robusta y considerada como significativa por expertos en arte. Contamos para ello con el amable permiso de la Fundación de Antoni Tàpies de Barcelona (Tàpies, 2001) para realizar la prueba sobre su colección de 434 imágenes digitalizadas de pintura y obra gráfica de Antoni Tàpies.

2. METODOLOGÍA

La metodología utilizada en nuestra investigación está basada en determinar características locales que produzcan una representación de la imagen versátil y sólida capaz de mostrar el contenido global y local al mismo tiempo, y que a su vez hagan robusta la descripción ante la oclusión parcial de objetos contenidos y la transformación de la propia 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. 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 este enfoque consiste en analizar las imágenes como un conjunto de regiones, describiendo solamente su apariencia e ignorando su estructura espacial. La representación BOW 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 BOW fue implementada por primera vez en el desarrollo de una sistema experto de reconocimiento de objetos (Willamowski, Arregui, Csurka, Dance & Fan 2004).

La construcción BOW requiere dos decisiones principales de diseño:

a. La elección de los descriptores locales que aplicamos en nuestras imágenes.
b. La elección del método que se utilice para obtener el vocabulario visual.

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

Esta representación de una imagen no contiene información acerca de las relaciones espaciales entre "palabras visuales", del mismo modo que la representación BOW mezcla la información relativa al orden de la palabras en los documentos. No obstante, los métodos BOW, 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.

Para superar las limitaciones del enfoque BOW se ha implementado la metodología PHOW (Pyramid histogram of visual words). La pirámide de coincidencias trabaja mediante la colocación de una secuencia de cuadrículas cada vez más finas sobre la imagen obteniendo una suma ponderada de la cantidad de coincidencias que ocurren en cada nivel de resolución de la pirámide.

Detallamos a continuación los pasos de la determinación.

2.1. Extracción de características locales

Lowe en el año 2000 describe un sistema de visión por computador para realizar el reconocimiento de objetos que también hace uso de las características locales de complejidad intermedia de la imagen que son invariantes a muchos parámetros. Los denomina descriptores SIFT (Scale Invariant Feature Transform); características invariantes a transformaciones de escala y con ellos consigue transformar una imagen en una representación que no se ve afectada por los cambios de escala y otras transformaciones similares.

Este proceso consigue la integración de las características de una manera similar al proceso de atención visual en serie que se ha demostrado que desempeña un papel importante en el reconocimiento de objetos en la visión humana.

El propio Lowe, en 2004, encuentra que la mejor solución de compromiso entre rendimiento y rapidez se obtiene usando una cuadrícula de muestreo de gradientes 16 x 16 y agrupando los histogramas en 4 x 4 (Fig. 1) en torno al punto de interés o *keypoint*. El descriptor final propuesto en esta formulación es 128 dimensional (4 x 4 x 8) (Fig. 2). Originalmente estos descriptores fueron desarrollados para el reconocimiento de objetos en general y para realizar la alineación de imágenes.

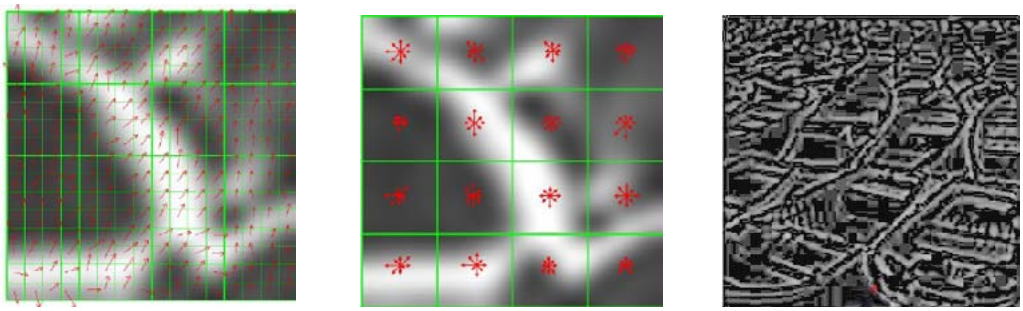

Fig.1. a) Keypoint. b) Región de 16 x 16 píxeles alrededor del keypoint y gradientes. c) Subregiones de 4 x 4 píxeles con histogramas de sólo 8 orientaciones.

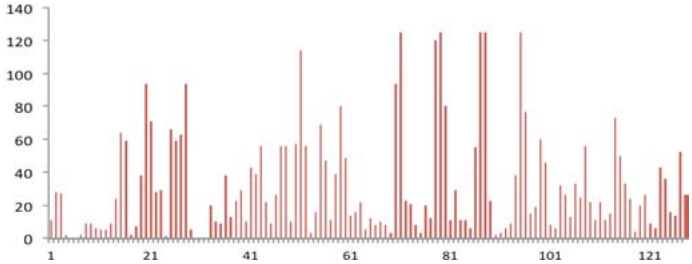

Fig. 2. Descriptor

Fig. 3. Imagen con una malla regular de 10 x 10 puntos de interés.

En algunos estudios (Lazebnik, Schmid & Ponce, 2006; Fei-Fei & Perona, 2005) el cálculo de los descriptores locales SIFT, en lugar de realizarse únicamente en los puntos de interés, se efectúa en los nodos de una malla regular sobreimpuesta en la imagen (Fig. 3). Este enfoque es preferible con el fin de mejorar la capacidad de discriminación en implementaciones orientadas a la clasificación de escenas, dado que, para determinados tipos de imágenes puede resultar poco representativo utilizer únicamente los puntos destacados. El punto de partida para la construcción n del vocabulario visual es el conjunto total de descriptores calculados para la colección de imágenes. y el objetivo que nos proponemos es obtener un vocabulario de "palabras visuales".

2.2. CONSTRUCCIÓN DEL VOCABULARIO VISUAL

La construcción del vocabulario se realiza mediante agrupación (clustering). Concretamente 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 euclidia ordinaria en los procesos de agrupación y cuantización, y elegimos el número de clústers dependiendo del tamaño deseado de vocabulario.

El algoritmo K-means busca la partición mediante la iteración de dos etapas. La primera etapa consiste en asignar cada descriptor al centroide más cercano. En la segunda etapa se recalculan los centroides de cada región, calculando el vector de medias de los descriptores que han sido asignados a cada región. En la Fig. 4 se describe, a modo de ejemplo 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 como se describe a continuación:

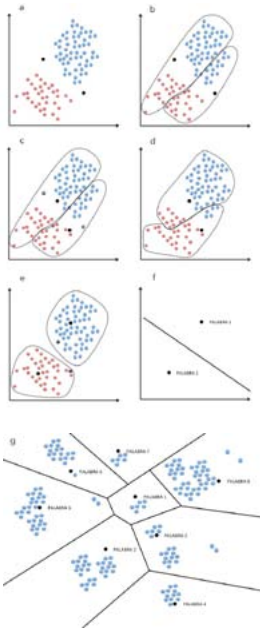

Fig.4. Algoritmo k-means

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 apreciable en los centroides.
g) Ilustra la partición del espacio de descriptores en el caso de un vocabulario de más palabras. Dado un descriptor determinado, calcularemos el centroide más cercano, y le corresponderá la palabra representada por dicho centroide.

De esta manera, dada una imagen con un conjunto de descriptores, podemos usar los centroides obtenidos en el algoritmo K-means para atribuir la "palabra visual" a la que pertenece cada descriptor buscando el centroide más próximo.

Para superar las limitaciones del enfoque de BoW, Lazebnik et al. (2006) proponen un método que incorpora con éxito información espacial al modelo BoW. Se denomina PHOW (Pyramid Histogram Of visual Words). En nuestro trabajo hemos implementado esta metodología de histogramas en pirámide que consiste en la colocación de una secuencia de rejillas cada vez más finas sobre la imagen, y en la obtención de una suma ponderada del número de "palabras visuales" coincidentes que se producen en cada nivel de resolución (L). Dada una resolución fija, se dice que dos puntos coinciden si están en el mismo cuadrante de la rejilla; las coincidencias encontradas en resoluciones más finas se ponderan más alto que las coincidencias encontradas en resoluciones más gruesas.

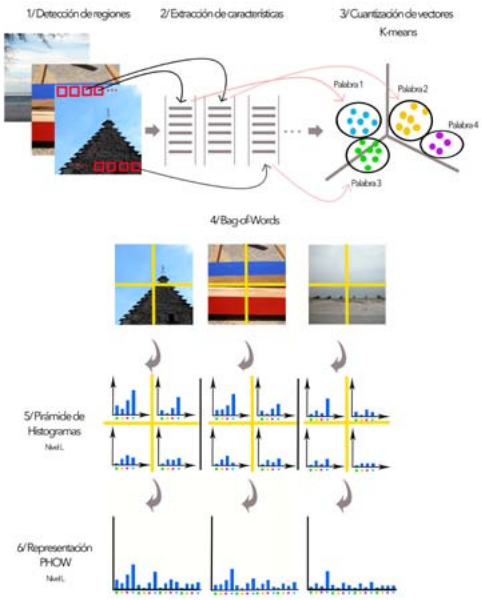

Fig. 5. Esquema del cálculo de la representación BOW y PHOW para una colección de imágenes.

En la Figura 5, se describe esquemáticamente la construcción del BOW a partir de las imágenes:
1- Detección automática de regiones/puntos de interés colocando una malla regular sobre la imagen.
2- Cálculo de descriptores locales sobre estas regiones.
3- Cuantizar los descriptores en palabras para formar el vocabulario visual mediante el algoritmo k-means.
4- Contabilizar las veces que ocurre en la imagen cada palabra específica del vocabulario con el objetivo de construir el BOW (histograma de palabras).
5- Cálculo de la Pirámide de Histogramas de L niveles.
6- Representación PHOW concatenando los histogramas de los diferentes niveles.

2.3. Representación de aspectos latentes

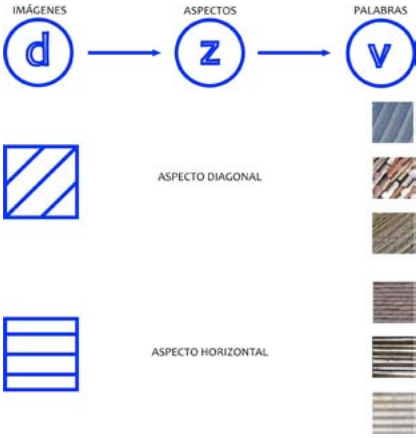

Fig. 6. Esquema del cálculo de la representación BOW y PHOW para una colección de imágenes.

El pLSA (Probabilistic Latent Semantic Analysis) es un modelo generativo que proviene del análisis estadístico de textos (Hofmann 2001). En este tipo de análisis de texto se utiliza para descubrir los temas de un documento mediante su representación como BOW. En nuestro caso, hay "imágenes" en lugar de "documentos" y en lugar de "temas" se descubren "categorías de objetos". De esta forma una imagen que contiene diferentes tipos de objetos se modela como una mezcla de temas.

Este modelo tiene la doble capacidad de generar una representación de escena bajo-dimensional robusta, y también de capturar automáticamente los aspectos significativos de la escena. 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 en un vocabulario visual establecido a partir de un proceso de cuantización como se ha señalado anteriormente. El método detectará en las imágenes categorías de objetos, patrones formales, de modo que una imagen que contiene varios tipos de objetos se modela como una mezcla de temas (Fig. 6).

3. RESULTADOS

Hemos aplicado el modelo descrito sobre colecciones de artista constituidas por imágenes digitales de escenas naturales de carácter abstracto (Rosado, Reverter, Figueras & Planas, 2014; Rosado, Figueras, & Reverter 2014) obteniendo categorizaciones significativas. En el presente estudio se da un paso más y se realiza la agrupación de imágenes de obra pictórica del artista Antoni Tàpies (Tàpies 2001) El reto al que nos enfrentamos, a diferencia de los estudios encontrados en la literatura que abordan la agrupación automática por contenido visual de imágenes de escenas reales y objetos cotidianos, es que aquellas tienen un contenido semántico universalmente asumido y en cambio, las bases de datos de arte abstracto que utilizamos en nuestro análisis son colecciones de imágenes de formas que el artista creador vincula porque considera que entre ellas existen analogías de sentido, y que por tanto suponen un reto de más difícil validación. Aclaramos que con el termino "abstracto" nos referimos al arte que no intenta imitar un modelo conocido, o sea, "no objetivo"

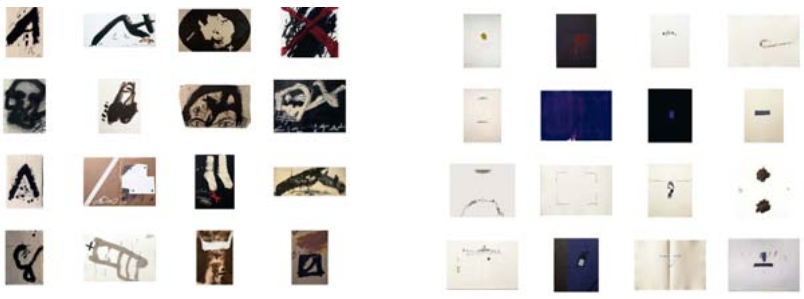

Fig.7. A la izquierda el aspecto: Trazo Grueso Denso y a la derecha el aspecto: Detalle sobre Fondo Plano. © Fundació Antoni Tàpies, Barcelona / Vegap. De la fotografia: © Gasull Fotografia.

En la Fig. 7 mostramos dos de las 13 categorías obtenidas en el análisis de 434 imágenes de esta colección, únicamente en función de su contenido visual.

En la Fig. 8 mostramos algunas de las 300 ""palabras visuales"" que conforman el vocabulario de la colección Tàpies y que han sido utilizadas para realizar las clasificaciones.

Fig.8. Palabras visuales constituidas por fragmentos de diferentes imágenes de la colección Tàpies

Cada una de las palabras muestran conjuntos de pequeñas regiones que se corresponden con la zona de 16 x 16 píxeles alrededor del keypoint que ha sido utilizado para calcular el descriptor SIFT de la zona. Estas pequeñas regiones de las imágenes, al ser visualizadas, contribuyen a la comprensión de las características formales del aspecto al que corresponden. Al mirarlas agrupadas se perciben las constantes que han motivado el agrupamiento en la misma "palabra visual". La posibilidad de visualizar el vocabulario particular que utiliza una artista plástico al ejecutar sus obras y a la vez de medir la frecuencia del uso de unas palabras sobre otras, resulta muy significativo y de utilidad para la comprensión y el estudio de su producción. A su vez, la posibilidad de configurar un vocabulario visual complejo más amplio, compuesto por palabras de diversos artistas, es muy sugerente y sería también de gran utilidad como fondo para la creación digital de nuevas posibilidades estéticas.

4. CONCLUSIONES

Los resultados obtenidos son considerados satisfactorios por expertos en arte y, lejos de pretender substituir el criterio de los entendidos, el sistema programado propone una herramienta de estudio para establecer analogías y buscar aspectos latentes en grandes colecciones de imágenes de arte abstracto, aunque también sería extensible el uso en obra figurativa. El sistema permite repetir los estudios sobre diferentes periodos del mismo artista, o sobre colecciones de distintos artistas o épocas, con los mismos criterios. De esta forma, los resultados obtenidos se pueden comparar sin riesgo de caer en interpretaciones subjetivas condicionadas por las preferencias o conocimientos previos.

Cabe destacar el interés de las herramientas presentadas desde el punto de vista del acceso simultáneo por parte de un artista a su colección de múltiples imágenes para poder analizar su trayectoria creativa, o desde el punto de vista de los teóricos del arte que podrían realizar estudios comparativos entre las obras de arte de diferentes artistas o épocas sin necesidad de mover de su emplazamiento ni una obra. Sin entrar a valorar la calidad estética de las agrupaciones que realiza la maquina, podemos concluir que las relaciones que establece, dada la cualidad matemática que le confiere la metodología utilizada para su realización, proporcionan nuevos puntos de vista libres de preconcepciones historicistas o vivenciales.

Referencias

Dondis, D.A. 1984. *La sintaxis de la imagen: introducción al alfabeto visual*. Barcelona: Gustavo Gili.

Fei-Fei, L. & Perona, P. 2005. "A Bayesian hierarchical model for learning natural scene categories". In

Proc. CVPR. San Diego, CA, USA.
Flusser, V. 2009. *Una filosofía de la fotografía*. Madrid: Síntesis.
Hofmann, T. 2001. "Unsupervised learning by probabilistic latent semantic analysis". *Machine Learning, 42*:177-196.
Lazebnik, S., Schmid, C. & Ponce, J. 2006. "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories". *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2.
Lowe, D.G. 2000. "Towards a Computational Model for Object Recognition in IT Cortex", en *First IEEE International Workshop, Biologically Motivated Computer Vision. BMCV 2000.* Berlín*:* Springer Berlin Heidelberg.
Lowe, D.G. 2004. "Distinctive Image Features from Scale Invariant Keypoints". *Int. Journal of Computer Vision,* 60 (2): 91-110.
Manovich, L. (2012). "¿Cómo ver 1000000 de imágenes?". *Deforma cultura on line*. [Accedido 11-01-2015]. <http://www.deforma.info/es/product.php?id_product=24>.
Rosado, P., Reverter, F., Figueras, E. & Planas, M.A. 2014. "Semantic-Based Image Analysis with the Goal of Assisting Artistic Creation". *Lecture Notes in Computer Science. 8671: 526-533.*
Rosado, P., Figueras, E. & Reverter, F. 2014. "Intersecciones entre visión artificial y mirada artística". *BRAC - Barcelona, Research, Art, Creation.* 2 (1): 1-54.
Tàpies, A. 2001. "Fundació Antoni Tàpies". [Accedido 12-01-2015]. <http://www.fundaciotapies.org/site/spip.php?rubrique65>
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.

Agradecimientos

[1] Agradecemos al Archivo de la Fundación Antoni Tàpies de Barcelona la posibilidad de acceder a la colección de imágenes digitales de obra del artista.