Anotación No Supervisada de Imágenes como Expansión Multimodal de Consultas

Luis Pellegrin, Hugo Jair Escalante

Reporte Técnico No. CCC-15-0003
27 de Agosto de 2015

Luis Enrique Erro 1
Sta. Ma. Tonantzintla,
72840, Puebla, México.

Anotación No Supervisada de Imágenes como Expansión Multimodal de Consultas

Luis Pellegrin Hugo Jair Escalante Balderas

Coordinación de Ciencias Computacionales
Instituto Nacional de Astrofísica, Óptica y Electrónica
Luis Enrique Erro # 1, Santa María Tonantzintla, Puebla, 72840, México
E-mail: {pellegrin,hugojair}@inaoep.mx

Abstract

La tarea de anotación automática de imágenes (automatic image annotation, AIA) consiste en asignar automáticamente palabras clave a imágenes, de tal manera que puedan describir su contenido visual. El interés en el desarrollo de sistemas en AIA ha aumentado en los últimos años, esto debido principalmente al crecimiento de colecciones de imágenes en la Web, las cuales demandan un correcto etiquetado para llevar a cabo tareas como la recuperación, categorización o exploración de imágenes.

Los métodos actuales para resolver la tarea de AIA trabajan bajo dos principales enfoques: supervisados y no supervisados. Ambos enfoques sufren de limitaciones en la anotación debido a la diversidad tanto visual como textual que dificultan la asignación correcta de palabras a las imágenes. Los métodos de anotación bajo el enfoque supervisado usan colecciones de imágenes etiquetadas, y se ven limitados a llevar a cabo la anotación de imágenes usando sólo las etiquetas presentes en la colección. En cambio, los métodos de anotación bajo el enfoque no supervisado utilizan una colección de imágenes de referencia, en donde para cada imagen se tiene un fragmento de texto asociado, del texto asociado se derivan las etiquetas para llevar a cabo la anotación de imágenes sin depender de un número fijo de etiquetas. Sin embargo, los métodos actuales de anotación de imágenes no supervisados requieren de estrategias eficaces para explotar características visuales y textuales.

Para sobrellevar limitaciones de asignación y explotar la información presentes tanto en imágenes como texto, se proponen nuevos métodos de anotación de imágenes bajo el estudio de AIA no supervisada en analogía a la expansión automática de consultas (automatic query expansion, AQE).

La analogía de AIA no supervisada con AQE nos permite definir un marco de trabajo que caracteriza a la tarea de anotación para: 1) explorar, analizar y proponer técnicas eficaces basadas en AQE para AIA no supervisada; 2) explotar y proponer esquemas de representación unimodales y multimodales, entre texto e imágenes, para llevar a cabo la anotación; y 3) analizar e identificar problemas durante el proceso de anotación bajo la analogía de AQE en donde se busca proponer soluciones.

Palabras clave: Anotación de imágenes, recuperación de imágenes, expansión de consultas, procesamiento de información multimodal, procesamiento de información multimedia.

Contents

1 Introducción

Los avances en desarrollo tecnológico, especialmente en dispositivos multimedia han facilitado la incorporación de imágenes, videos, audios, entre otros medios a Internet. En términos de imágenes: *'el almacenamiento de bajo costo y el fácil manejo de servidores Web han alimentado la metamorfosis del hombre común de un consumidor pasivo de fotografías (en el pasado) a un productor activo hoy en día'* [23]. Consecuentemente, los datos multimedia, no sólo imágenes (documentos de texto, audios y videos), se han incrementado considerablemente. Existen grandes colecciones de imágenes compartidas y mantenidas por usuarios en la Web, éstas pueden ser colecciones de dominios específicos (por ejemplo aviones[1]), redes sociales (*Facebook*[2]*)*, sistemas Web para almacenar y compartir archivos multimedia (*Flickr*[3]*, Picasa*[4]), así como diversos *blogs*. En las colecciones de imágenes de redes sociales y sitios Web para almacenar y compartir archivos multimedia los usuarios llevan a cabo su propia anotación, es decir, los usuarios suelen asociar un texto descriptivo (llamado comúnmente etiqueta) a las imágenes, videos, audios, entre otros medios. Sin embargo, algunas de las etiquetas proporcionadas por los usuarios no suelen describir fielmente la información multimedia contenida. Paralelamente al incremento de información multimedia ha surgido una gran demanda por medios efectivos y eficientes para organizar e indexar datos, de tal manera que se pueda recuperar información útil cuando sea requerida [63].

Una estrategia inicial para llevar a cabo la tarea de anotación de imágenes consiste en realizarlo manualmente, lo que puede resultar impráctico al trabajar con enormes cantidades de imágenes (un estudio sobre diferentes estrategias eficaces para la creación de etiquetas para anotación puede ser consultado en [36]). El *crowdsourcing* es una bien conocida estrategia de anotación manual de forma colectiva, en donde la dificultad reside en asegurar que un grupo de no expertos realicen el etiquetado de imágenes de manera precisa, manteniendo un proceso rápido y a la vez económico [83]. Los procesos de *crowdsourcing* a pesar de ser colaboraciones eficientes tienen que pasar por un filtro de calidad donde se evalúa el trabajo individual de los usuarios involucrados en la anotación, y como resultado sólo se obtiene un porcentaje de anotaciones confiables en las imágenes. Debido a lo anterior, existe un interés creciente en el desarrollo de métodos automáticos para llevar a cabo la anotación de imágenes.

La tarea de anotación automática de imágenes (*Automatic Image Annotation, AIA*), se encarga de encontrar una descripción textual que defina el contenido visual en las imágenes. Los sistemas de AIA requieren del desarrollo de sistemas que dada una imagen de entrada, se lleve a cabo un proceso que consiste en asociar etiquetas (llamadas también conceptos) a imágenes, ésto buscando describir de manera coherente el contenido visual de la imagen [58, 89]. La principal razón para llevar a cabo la tarea de AIA es para simplificar el acceso a las imágenes [36]. Una de las consecuencias directas de lograr una anotación confiable con sistemas de AIA es que la búsqueda de imágenes basada en consultas textuales puede tener mayor significado semántico [23].

Actualmente, las consultas basadas en texto son ampliamente usadas por la mayoría de los usuarios para buscar imágenes en la Web [48, 62]. Bajo este paradigma de consulta, si el número de palabras clave usadas en la consulta son pocas (por ejemplo una o dos palabras), la descripción de la consulta puede llegar a ser muy general, llegando a caer en un problema de ambigüedad. Por ejemplo, considerando una consulta formada por la palabra 'jaguar', esta palabra puede referirse a un automóvil, un felino, u otras definiciones bajo diferentes contextos (ver Figura 1).

[1] http://airliners.net

[2] https://www.facebook.com/

[3] https://www.flickr.com/

[4] http://picasa.google.com/

Figure 1. Imágenes recuperadas en un motor de búsqueda utilizando como palabra clave 'jaguar'.

Las consultas cortas si no son tratadas correctamente pueden llevar a resultados de búsqueda desfavorables. Una de las razones inherentes a este problema es que la mayoría de los motores de búsqueda actuales en Internet explotan eficientemente el texto para recuperar imágenes, pero ignoran el contenido de la imagen [58], es decir, y nuevamente reiterando, requieren de una anotación confiable en las imágenes.

La anotación de imágenes no es una tarea trivial y presenta varias dificultades, como se indica a continuación. En AIA se suele trabajar con una amplia variedad visual [69], lo que conlleva a considerar múltiples conceptos a anotar, y existe un problema de brecha semántica (*semantic gap*) definido como: *'...la falta de coincidencia entre la información que puede ser extraída de los datos visuales y la interpretación que tiene un usuario en una situación dada para estos datos'* [82]. Este problema de falta de correspondencia ha surgido en numerosas investigaciones, en [37] se caracteriza el fenómeno a través una jerarquía en niveles del procesamiento llevado a cabo para anotar una imagen: 1) datos crudos, representando a la imagen; 2) descriptores visuales, vectores numéricos que representan características visuales de la imagen; 3) objeto, combinaciones de descriptores visuales; 4) etiquetado de objectos, nombres simbólicos dados a los objectos; y 5) semántica, relaciones entre los objectos (ver Figura 2).

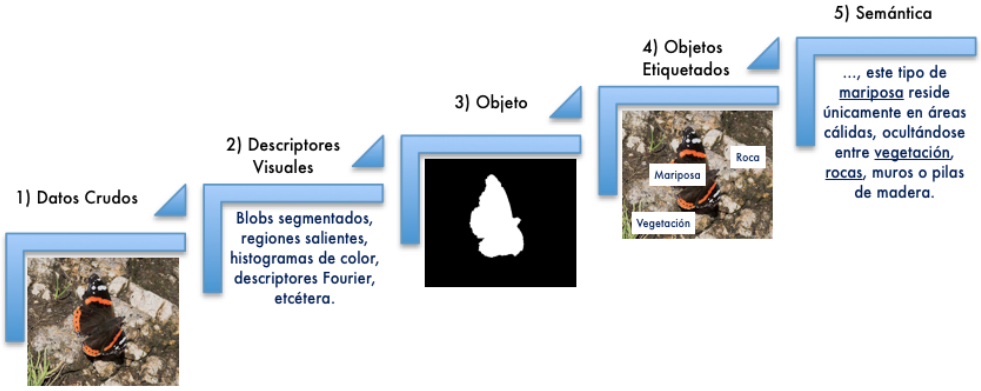

Figure 2. Jerarquía en niveles de procesamiento para llevar a cabo la anotación de una imagen: desde datos crudos hasta semántica (Figura adaptada de [37]).

A través de la caracterización de [37] se puede notar que automatizar una posible interpretación visual para llevar a cabo la anotación de las imágenes debe pasar por varios procesos dificultando la tarea.

Dos problemas inherentes a la brecha semántica son la polisemia visual y la sinonimia visual. La polisemia visual se presenta cuando imágenes con alta similitud visual (la cual se asume por la distancia muy cercana en sus descriptores visuales), presentan en su contenido visual diferentes significados semánticos [87]. La polisemia visual suele presentarse con mayor frecuencia al trabajar con pequeñas colecciones de imágenes y trae como consecuencia una baja precisión en la anotación [87] (por ejemplo: *'no todo lo que es redondo y rojo representa a una pelota'*, ver Figura 3).

Figure 3. Polisemia visual, imágenes con contenido visual similar a una pelota que representan diferentes contenidos semánticos.

En cambio, en la sinonimia visual (Figura 4), es el caso de diferentes imágenes que describen semánticamente el mismo significado (objeto), sin embargo, no son imágenes con alta similitud visual, es decir, no son cercanas en distancia de sus descriptores visuales [87].

Figure 4. Sinonimia visual, una bicicleta se representa por diferentes imágenes que comparten un mismo significado semántico. Imágenes tomadas de la BD en [93].

Una de las razones de que la polisemia y sinonimia visual continúan como dificultades presentes en el problema de brecha semántica se debe a que los descriptores visuales extraídos de la imagen y usados para la representación de su contenido, como *SIFT* (*Scale Invariant Feature Transform*), aún son sensibles a

variaciones como iluminación, traslación o distorsión en las imágenes [89].

Con el objetivo de dar solución o acortar el problema de brecha semántica, se han propuesto técnicas de visión por computadora que permiten la extracción automática de características de bajo nivel de las imágenes con cierto grado de eficacia [49], sin embargo, la problemática en la tarea de AIA aún reside en cómo derivar conceptos de alto nivel automáticamente del contenido visual y de la información disponible [49].

Los métodos actuales para resolver la tarea de AIA trabajan bajo dos principales enfoques: supervisados y no supervisados. En el enfoque de AIA supervisada, se trabaja usando una colección de imágenes etiquetadas y sólo se anotan imágenes con las etiquetas presentes dentro de la colección limitando la escalabilidad de anotación. En cambio, en el enfoque de AIA no supervisado, las investigaciones iniciales han buscado mayor diversidad de etiquetas para la anotación y no usan una colección de imágenes anotadas. En AIA no supervisada se usa una colección de imágenes de referencia, en donde para cada imagen de la colección se tiene un fragmento de texto asociado [93, 29, 24]. En el escenario de AIA no supervisada cualquier palabra extraída del texto asociado en la colección puede usarse como etiqueta.

El proceso de anotación general usado en AIA no supervisada usa dos pasos principales: 1) una recuperación de k imágenes similares visualmente a la que se desea anotar, y 2) un proceso de minería de texto aplicado en el texto asociado de las k imágenes para derivar las etiquetas a anotar. En el esquema de AIA no supervisada los pasos son procesados en secuencia, por lo tanto se tratan las características visuales y textuales por separado, por lo cual no explotan adecuadamente las relaciones latentes entre los dos tipos de datos. La presente investigación está enfocada en proponer distintas alternativas para tomar mayor ventaja de las relaciones entre características visuales y textuales buscando mejorar el rendimiento en compsaración a los métodos actuales en AIA no supervisada. Hemos definido una analogía entre la AIA no supervisada y la expansión automática de consultas, de la cual proponemos un marco de trabajo para evaluar las distintas estrategias y sacar provecho de la información. Así mismo, proponemos entre estas estrategias el uso de información multimodal (que proviene de dos o mas modalidades, para este caso texto e imágenes) bajo dos diferentes paradigmas de anotación de imágenes, local y global, que hemos definido a través de la analogía que proponemos. Además, proponemos el desarrollo de una métrica que permita estimar la complejidad de la imagen a anotar, el objetivo es dada una imagen y una colección de imágenes de referencia identificar bajo que paradigma de anotación de imágenes es posible sacar el mayor provecho.

El reporte es estructurado como sigue. En la Sección 2, se presentan los fundamentos a expansión automática de la consulta, tema que usamos como analogía para desarrollo de la investigación. En la Sección 3 se presenta el trabajo relacionado, exponiendo los diferentes enfoques en los que se lleva a cabo el proceso de anotación de imágenes. La Sección 4 incluye la justificación y motivación de la investigación. En la Sección 5, la propuesta de investigación es presentada, en donde primero proponemos una definición de AIA no supervisada en analogía a AQE y definimos el marco de trabajo, después se presentan las preguntas de investigación, los objetivos y la metodogía propuesta para llevar a cabo la investigación. En la Sección 6, se presenta el diseño experimental que se llevará a cabo en la metodología. En la Sección 7 se presentan los resultados preliminares. Finalmente, la Sección 8 incluye las conclusiones, el plan de publicaciones y el cronograma de trabajo.

2 Fundamentos

En esta sección se describe las bases teóricas referentes a expansión automática de consultas que servirá como tema base para la propuesta de investigación.

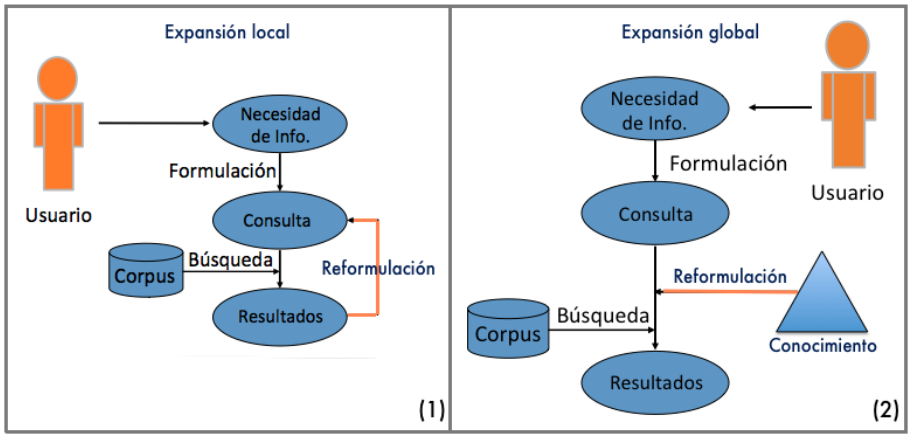

Figure 5. Dos de los principales paradigmas para llevar a cabo la expansión: (1) expansión local, después de obtener resultados se reformula la consulta expandiendo con términos extraídos de los mismos resultados; (2) expansión global, inmediatamente después de que es formulada la consulta es expandida a través de una reformulación, el conocimiento global puede ser información estadística de una corpus, un thesaurus, o conocimiento obtenido por algún método de aprendizaje.

2.1 Expansión automática de consultas

En el área de recuperación de información (*information retrieval*, IR), la expansión automática de consultas (*Automatic Query Expansion, AQE*) consiste en expandir la consulta del usuario agregando palabras que aumentan la descripción de la información que el usuario requiere. La utilización de AQE es una forma bien conocida de sobrellevar el problema de falta de correspondencia en IR: '*la inefectividad de los sistemas de IR se debe en gran medida a la inexactitud, con la cual, la consulta es formada por pocas palabras clave que modelan la información que el usuario necesita*' [16].

Las técnicas de AQE pueden ser clasificadas en cinco grupos de acuerdo al paradigma conceptual usado para encontrar la expansión [16]:

1. Específicas a la consulta local. Este paradigma es conocido como expansión local (Figura 5(1)), éste usa técnicas que toman ventaja de la información local que es recuperada a través de la consulta para llevar a cabo la expansión.

2. Basadas en el corpus. Éste paradigma es conocido como expansión global (Figura 5(2)). Las técnicas realizadas bajo este esquema analizan el contenido de la colección o corpus de entrenamiento con el objetivo de identificar características útiles para llevar a cabo la expansión.

3. Análisis lingüístico. Se usan técnicas que explotan propiedades del lenguaje para formar relaciones que sirvan para expandir o reformular la consulta.

4. Análisis de bitácora de búsqueda (*search log analysis*). Incluye técnicas que analizan el hábito de los usuarios en registros de bitácoras, con la finalidad de crear asociaciones útiles que puedan ser sugeridas durante la consulta.

5. Datos Web. Bajo este paradigma se usan herramientas que analizan relaciones entre los enlaces de la Web con la finalidad de adquirir información útil para usar en la expansión.

Para algunos casos la expansión de la consulta puede llevarse a cabo de forma iterativa, es decir, se realiza la expansión mediante algún paradigma y posteriormente se reajustan y agregan nuevos términos a la consulta y se vuelve a lanzar la consulta. Con este proceso iterativo se busca reforzar la decisión de la consulta tratando de mejorar la correspondencia entre la consulta y la intención del usuario para llevar a cabo. Al proceso iterativo de reformular la consulta se le conoce como *pseudo relevance feedback* [16].

Aplicaciones de AQE han tenido un gran desarrollo en los últimos 20 años [16], creando diversas técnicas que le han permitido a AQE trabajar con una amplia gama de fuentes de información y han ayudando a incrementar la eficiencia computacional. La madurez alcanzada en AQE le ha permitido el desarrollo de métodos para tomar en cuenta dependencia de términos, expansión de consultas estructuradas, modelos de interacción, aprendizaje adaptativo y posibilidades de híbridos [16]. La diversidad de métodos y la analogía con AIA no supervisada que hemos definido en este reporte (en las subsecciones 5.1 y 5.2), nos lleva a argumentar que métodos de AQE pueden ser utilizados para sacar provecho de la información textual y visual, y para proponer nuevos de métodos para AIA no supervisada.

3 Trabajo relacionado

El proceso de anotación de imágenes se ha llevado a cabo mediante tres enfoques generales (Tabla 1). Bajo un enfoque supervisado de anotación de imágenes se usa una colección de imágenes etiquetadas, mientras que bajo el enfoque no supervisado generalmente se usa una colección de imágenes de referencia en la que las imágenes no están etiquetadas. El enfoque semi-supervisado está ubicado de forma intermedia entre estos enfoques y puede ser visto como un caso de anotación supervisada en el que sólo se cuenta con un subconjunto de imágenes etiquetadas de la colección.

Enfoque	**Colección de imágenes**	**Método general de anotación**
Supervisado	etiquetadas	aprendizaje de correspondencia entre etiqueta(s) e imágenes
Semi-Supervisado	un subconjunto etiquetado	propagación de etiquetas
No Supervisado	sin etiquetar	extracción de etiquetas del texto asociado a las imágenes

Table 1. Enfoques generales de AIA

3.1 AIA Supervisada

En la AIA supervisada se cuenta con un colección de imágenes etiquetadas. La colección está compuesta por un conjunto de ejemplos, en el que cada ejemplo consiste de un vector de características visuales extraídas de la imagen y las etiquetas asignadas como anotación. La colección de imágenes con sus respectivas etiquetas se utilizan para aprender correspondencias entre las características visuales y las etiquetas. Generalmente se pueden identificar dos tipos de modelos de aprendizaje usados para la fase de entrenamiento [75]:

- Discriminativos. Los modelos discriminativos modelan la distribución de probabilidad condicional $p(y|X)$, en donde las variables de entrada X son vectores de características visuales extraídas de las imágenes, y la variable de salida y es la etiqueta a anotar.

- Generativos. En el modelo discriminativo generalmente se provee un modelo por cada etiqueta de anotación, mientras que en un modelo generativo se provee un modelo probabilístico completo de todas las variables, es decir, se modela una función de densidad de la probabilidad condicional $p(X, Y)$ en donde Y son las etiquetas a anotar.

3.1.1 Modelos discriminativos

Las máquinas de soporte vectorial (*support vector machine*, SVM) [15, 22, 94, 45], k vecinos más cercanos (*k-nearest neighbour*, KNN) [38, 42], redes neuronales artificiales y árboles de decisión son algunos de los algoritmos que se utilizan para generar o como modelos discriminativos en la AIA supervisada.

En [15, 22, 94, 45] se usa SVM para llevar a cabo la tarea de anotación de imágenes generando modelos discriminativos usando estrategias *one-versus-all* (*OVA*) [15, 22, 45] y *one-versus-one* (*OVO*) [94]. Para la estrategia *OVA* por cada etiqueta a anotar se entrena una SVM binaria para aprender un modelo discriminativo contra las demás etiquetas a anotar, generando n clasificadores. En cambio, para la estrategia *OVO* por cada par de etiquetas a anotar se entrena un SVM, generando $n(n-1)/2$ clasificadores. Generar modelos discriminativos para anotación de imágenes por estrategias *OVA* o *OVO* usando SVM generalmente requiere alto procesamiento computacional.

El uso de KNN como modelo no requiere entrenar n clasificadores. En [38] se usa una herramienta de anotación que permite segmentar la imagen y etiquetar manualmente formando así un diccionario de imágenes. Cuando se tienen suficientes imágenes en el diccionario, dada una nueva imagen sin anotación, la herramienta usa vecinos más cercanos para anotarla de acuerdo a su similitud con imágenes en el diccionario. Sin embargo, la herramienta de [38] requiere procesar la similitud de la imagen a anotar contra todas las imágenes en el diccionario. Una estrategia para disminuir el número de comparaciones de similitudes a calcular, es generar prototipos, es decir, construir elementos representativos que generalicen a subconjuntos de imágenes, reduciendo el número de elementos a comparar. En [42] se lleva a cabo anotaciones de diferentes colores a las regiones de las imágenes, en donde previamente de la colección de imágenes se generan prototipos por cada color por medio de centroides realizando *clustering* en la representación RGB de las imágenes. Sin embargo, es difícil generar prototipos para colecciones de imágenes que presentan múltiples etiquetas anotadas por imagen y una gran variedad de diversidad visual en las imágenes.

Cabe señalar que para la construcción de modelos discriminativos se requiere de una colección de imágenes etiquetadas que generalmente se anotan manualmente, en donde la anotación puede llegar a ser subjetiva. En experimentos realizados con *Corel Dataset* [25] se ha mostrado más del 50% de palabras asignadas a tópicos no ocurren en la correspondiente imagen [9].

3.1.2 Modelos generativos

La AIA supervisada que utiliza modelos generativos ha sido un área activa en la última década. Se pueden identificar dos grupos principales de métodos en los que predomina el uso de probabilidades para generar modelos:

- Aquellos que usan ocurrencias y co-ocurrencias entre etiquetas y características visuales para estimar distribuciones de probabilidad conjunta [43, 51, 96]. Incluyendo métodos que usan técnicas estadísticas para capturar la máxima entropía entre las características visuales y textuales [44].

- Otro grupo que modela tópicos (LDA (*latent Dirichlet allocation*), LSA (*latent semantic analysis*), pLSA (*probabilistic LSA*)), donde cada tópico representa una distribución de probabilidad entre etiquetas y características visuales [5, 6, 17, 71, 46, 64, 95, 13, 14].

Los modelos generativos usados para AIA han sido criticados por varias razones: (1) maximizan la verosimilitud de los datos, lo cual no es del todo correcto para llevar a cabo la predicción de etiquetas, debido a que no todas las etiquetas asignadas a un tópico ocurren en su correspondiente imagen [71], (2) hacen suposiciones sobre la distribución de los datos para hacer tratable el aprendizaje del modelo, (3) algunos modelos generativos requieren de complejas inferencias [46], (4) se requieren parámetros adicionales a aprender debido a suposiciones en la distribución de los datos [71], (5) alto consumo de recursos tanto en procesamiento como en memoria, y (6) requieren de imágenes etiquetadas, en donde la anotación de grandes cantidades de imágenes resulta costosa debido a que se lleva a cabo manualmente.

Métodos que han surgido para subsanar problemas de rendimiento son [95] y [91]. En [95] se define un modelo de representación que puede ser visto como un modelo de espacio semántico en el que los tópicos son determinados por similitudes entre la cantidad de palabras y la distribución multivariada de Gausianas de las características visuales. Con la definición propuesta por [95] se persigue relajar las relaciones entre modalidades sin caer en independencias en el modelado de tópicos. En [13, 14], se define una representación multimodal que combina características visuales y texto mediante el uso de factorización de matrices, la idea es encontrar factores latentes que correlacionen los datos multimodales en el espacio de representación de la matriz. Una extensión al modelo de [13, 14] se realiza en [91] donde la ventaja radica en que el aprendizaje es realizado *online* mediante una formulación de descenso de gradiente.

El uso de *deep learning* (aprendizaje profundo) ha sido una tendencia en *machine learning* (aprendizaje automático) para obtener modelos generativos. *Deep learning* es inspirado y trata de replicar en computadora la arquitectura de procesamiento del cerebro humano, en la cual se cree que existe una serie de niveles de procesamiento con diferentes niveles de abstracción [65]. Una de las principales críticas en el área de *deep learning* es que los modelos generados tienen millones de parámetros y requieren grandes cantidades de datos etiquetados para el entrenamiento, los cuales son difíciles de obtener [65, 8]. Una de las principales aplicaciones para *deep learning* ha sido el aprendizaje de representaciones [8] que incluyen *deep auto-encoders* que consisten en métodos para reducir la información de diferentes modalidades generando un modelo multimodal [68, 84, 77].

3.1.3 Observaciones de AIA supervisada

Una limitante a resaltar de la AIA supervisada es que sólo es posible anotar con las etiquetas de la colección de imágenes que se use, las cuales suelen ser un número finito y pequeño que limitan las posibilidades de escalabilidad en la anotación. Usando métodos para generar modelos discriminativos se tiene la desventaja que es necesario entrenar un clasificador por cada etiqueta a anotar aumentando el procesamiento de computo, además, cabe señalar que algunos algoritmos de clasificación como es el caso de SVM es necesario proveer para el aprendizaje ejemplos positivos y negativos por cada etiqueta a anotar.

Para la AIA supervisada que usa modelos generativos no es la excepción el aprendizaje finito de etiquetas a aprender. Una de las bases de datos que ha sido ampliamente usada es el *Corel Dataset* [25] para evaluar modelos generativos para AIA supervisada. Diversos trabajos reportan resultados sobre esta BD [43, 44, 17, 6, 5, 13, 14, 64, 71, 96] en donde se usan 5,000 imágenes de 50 CDs, en donde cada CD representa corresponde a una etiqueta, es decir, sólo se consiguen anotar 50 etiquetas diferentes entrenando con 5,000 imágenes, lo cual hace resaltar que: 1) se requieren de muchos ejemplos de entrenamiento, e incrementa el número de ejemplos a medida que se agregan nuevas etiquetas, lo cual conlleva a mayor costo computacional; 2) resulta costoso construir BDs para entrenamiento para anotación de imágenes debido a que se requieren un número considerable de etiquetas y ejemplos para cada una de ellas, lo cual se realiza mediante anotaciones manuales (cabe señalar que debido a esto existen pocas BDs para evaluación).

3.2 AIA Semi-Supervisada

El enfoque de AIA semi-supervisado puede ser visto como un caso especial del enfoque supervisado. La diferencia con el enfoque supervisado radica en que en la colección de imágenes no se tienen etiquetadas todas las imágenes, sólo un subconjunto de ellas. La tarea de AIA semi-supervisada puede llevarse a cabo aprendiendo modelos del conjunto etiquetado, y después usar el conjunto no etiquetado como prueba e ir agregando los nuevos ejemplos aprendidos [35], en este caso se asume que los ejemplos aprendidos que se agregan han sido correctamente anotados. En [60] se usa un ensamble de clasificadores y se considera durante el entrenamiento a las imágenes sin etiquetas. El ensamble funciona en forma de cascada, en donde las imágenes sin etiqueta son clasificadas y sirven de entrada a un siguiente clasificador que considera la clasificación anterior.

Otra forma de ver la tarea de AIA semi-supervisada es como una propagación de etiquetas a imágenes sin anotación. En [4] las propagaciones se realizan localmente entre etiquetas, en donde para realizar la propagación de etiquetas es necesario establecer un parámetro n que determina la cantidad de etiquetas a propagar. En cambio en [19, 57] se consideran propagaciones holísticas, la limitante es que no es posible separar las similitudes entre diferentes etiquetas [4]. La suposición principal de [57] es que dos clases de etiquetas tienden a tener traslape en sus datos si comparten alta similitud. Otro trabajo en propagación es [80], se representan a las imágenes como nodos en un grafo que conectan por similitud visual o por co-ocurrencia de etiquetas, y la propagación se realiza en el grafo a imágenes sin etiqueta.

3.2.1 Observaciones de AIA semi-supervisada

La AIA semi-supervisada es un enfoque económico para anotar imágenes cuando se cuenta sólo con un conjunto reducido de imágenes anotadas [4]. Los trabajos de AIA semi-supervisada [4, 35, 19, 57] son métodos robustos a la anotación de imágenes que no necesitan contar con la anotación completa en la colección de imágenes. Sin embargo, permanece la problemática de que sólo es posible anotar con las etiquetas que se cuentan en la colección y por lo tanto se pierden diversidad de etiquetas en la anotación de imágenes.

3.3 AIA No Supervisada

La tarea de AIA no supervisada tradicional se describe de la siguiente manera. Dada una imagen a anotar y una colección de imágenes de referencia, en donde cada imagen en la colección no está etiquetada pero cuenta con un fragmento de texto asociado, la meta de la tarea consiste en asignar palabras que describan el contenido visual de la imagen a anotar empleando el texto asociado de las imágenes de la colección de referencia.

Investigaciones actuales [58, 93, 56, 78, 90, 74], han llevado a cabo la tarea de AIA no supervisada de la siguiente manera. Dada una imagen a anotar y una colección de imágenes de referencia, se encuentran k imágenes vecinas más cercanas visualmente de la colección. Posteriormente se procesa la información textual asociada de las k imágenes para llevar a cabo la anotación. Bajo este esquema de anotación se puede identificar dos pasos o módulos principales (ver Figura 6): (1) La recuperación de las k imágenes similares de la colección de imágenes de referencia, usando un módulo de recuperación de imágenes basado en contenido (*content-based image retrieval, CBIR*); (2) un módulo de minería de texto para el texto asociado de las k imágenes, de donde se produce la anotación a la imagen. Los pasos antes mencionados son usados de manera secuencial, en donde se identifica que la importancia de la anotación recae en el contenido visual de k imágenes, lo cual en ciertos casos puede no ser lo adecuado.

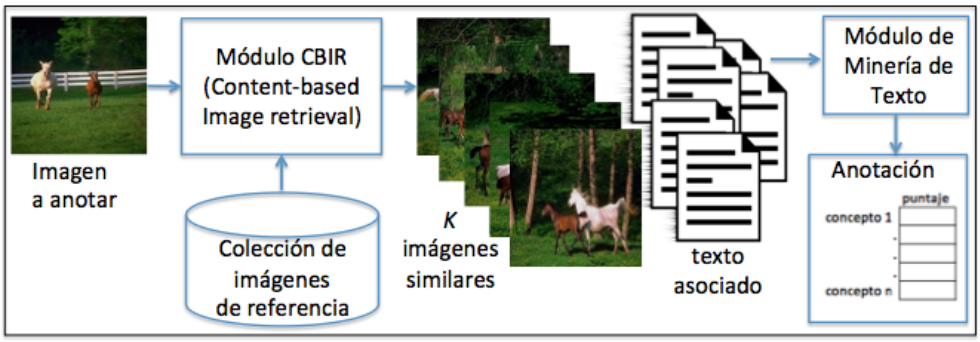

Figure 6. Enfoque tradicional de AIA no supervisada.

Desde un punto de vista práctico, en el enfoque de AIA supervisada sólo se anotan las etiquetas que se encuentran dentro de la colección de imágenes que se use. En cambio, el enfoque de AIA no supervisada en lugar de contar con etiquetas anotadas en las imágenes, se usa para derivar las etiquetas a anotar el texto asociado a imágenes de una colección de imágenes de referencia, en donde generalmente se cuenta con un gran número de imágenes y por lo tanto la cantidad de etiquetas a elegir aumenta. El texto asociado a la imagen generalmente es no estructurado y puede presentarse en diferentes formas: i) texto que acompaña a la imagen como puede ser el de artículos científicos, revistas, libros, entre otras fuentes de datos; ii) texto en las páginas Web en donde fue encontrada la imagen, incluyendo títulos, URL, subtítulos, entre otras características; iii) texto obtenido de un motor de búsqueda, como el de las palabras clave empleadas, *snippets*[5], información procesada del motor de búsqueda (por ejemplo el *ranking*), etcétera.

Bajo el escenario de AIA no supervisada es posible conseguir mayor diversidad en la anotación a diferencia del enfoque supervisado. Algunas de las razones para considerar al texto asociado para extraer las etiquetas a anotar son [49]:

- El texto codifica mejor la semántica de alto nivel que el contenido de la imagen.
- Muchas palabras y expresiones aparecen junto a la imagen y pueden ser relacionadas con esta.

En la AIA no supervisada, nótese que aunque no se cuenta con un etiquetado por imagen se debe contar con un texto asociado a cada imagen en la colección de imágenes de referencia. Sin embargo, no entra en contradicción con la definición de no supervisado que es aplicado cuando no se cuenta con conocimiento a priori que nos diga la salida deseada a cada entrada [65], pues en el caso de AIA no supervisada el texto asociado a cada imagen es generalmente no estructurado y no nos dice directamente qué anotar a la imagen complicando la tarea.

En AIA no supervisada es deseable que las etiquetas a anotar describan el contenido visual de la imagen. Sin embargo, la interpretación del contenido visual de una imagen es subjetivo y difícil de enseñar a una computadora por lo que el número de etiquetas que pueden ser anotadas es desconocido. Intuitivamente, elegir pocas etiquetas (una o dos) podría sugerir una anotación tomando en cuenta el contenido visual general o predominante, en cambio, un número de etiquetas (por ejemplo mayor a cinco) sugiere que la anotación tiene un mayor grado de especificidad.

Existen dos subproblemas diferentes en AIA no supervisada que se pueden caracterizan por: dar mayor importancia al contenido textual ó el enfoque tradicional que da mayor importancia al contenido visual.

[5]Pequeñas descripciones que aparecen en los resultados de los motores de búsqueda.

Cuando se da mayor importancia al procesamiento textual [20, 52, 24, 29]. Básicamente, la meta de la tarea es dada una imagen y su texto asociado, se extraen etiquetas del texto asociado que describan el contenido visual de la imagen (ver Figura 7), como ayuda se hace uso de herramientas de la Web para identificar la importancia y expresividad de la palabras (por ejemplo WordNet [20], *flickr* [52, 24]). La dificultad de anotación bajo este escenario es filtrar etiquetas irrelevantes y elegir a las que serán anotadas a la imagen. Nótese que es necesario contar con un texto asociado para la imagen a anotar, y no es necesario un paso de CBIR pues se ignora completamente el contenido visual de las imágenes.

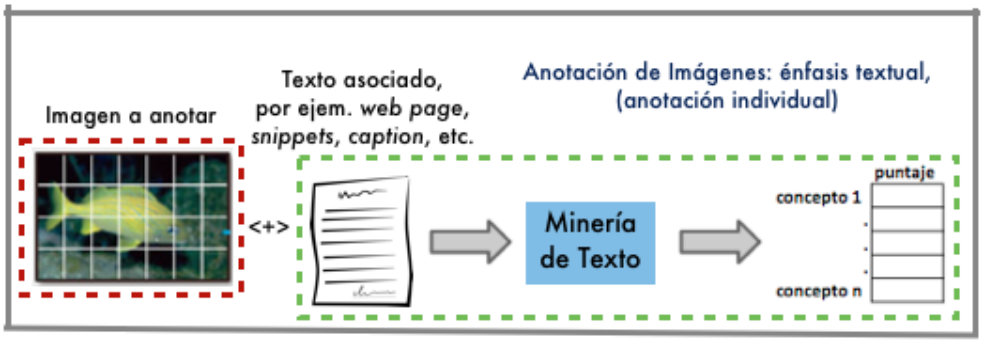

Figure 7. Proceso de anotación de imágenes individual. La imagen a anotar es anotada usando el texto asociada a ella ignorando por completo el contenido visual de la imagen.

En cambio, en el enfoque tradicional de AIA no supervisada la tarea se centra en etiquetar imágenes que no tienen ningún texto asociado, pero se toma en cuenta el texto asociado de una colección de imágenes de referencia. Trabajos que realizan la anotación no supervisada tradicional [58, 93, 56, 78, 90, 74], proponen variantes en el paso de recuperación y diferentes estrategias para el paso de minería de texto. En [58] para el paso minería de texto se usa una anotación con estrategia voraz (*greedy*) para ir eligiendo cada etiqueta a anotar, en donde se eligen las etiquetas que presentan mayor frecuencia en el texto asociado de la primera imagen recuperada en el paso de CBIR, en base a esta primera recuperación se toma en cuenta al texto asociado de las siguientes k imágenes. En [56] se emplea como colección de imágenes de referencia un subconjunto de imágenes de *Flickr*, y se propone para el paso de minería de texto usar la información de usuarios para asignar relevancia a las diferentes etiquetas a anotar, el supuesto es que las personas etiquetan a las imágenes visualmente similares con las mismas etiquetas.

En [74] para el paso de anotación se utiliza la técnica de pesado BM25 (*Okapi best matching 25*) para asignar un puntaje al conjunto de etiquetas extraídas del texto asociado anotando aquellas etiquetas de mayor puntaje. En [90] se propone una medida de estimación de relevancia para decidir las etiquetas a anotar. La medida de relevancia de [90] toma en cuenta la distribución de frecuencia de la etiqueta a evaluar en el texto asociado de las k imágenes y la distribución de frecuencia dentro de la misma etiqueta dentro de la colección de imágenes de referencia.

El método de anotación en [93] usa un análisis estadístico de coocurrencias de los términos del texto asociado de las imágenes de la colección de referencia. Primero se lleva a cabo un paso de CBIR para obtener k imágenes cercanas visualmente a la imagen que se desea anotar. Después, para el paso de minería de texto usa el texto asociado a las k imágenes recuperadas y calcula un puntaje para cada etiqueta posible a anotar (de una lista de n conceptos posibles), este puntaje es calculado usando la probabilidad de co-ocurrencia de las etiquetas con los otros términos presentes en el texto asociado a las k imágenes. Un

aspecto interesante del trabajo en [93] es que aunque lleva a cabo la anotación a dos pasos, el análisis de co-ocurrencias sobre los términos del texto asociado se hace sobre la colección imágenes de referencia completa. Una extensión al trabajo de [93] se realiza en [78], en donde se toma en cuenta una configuración heterogénea de descriptores visuales para el paso de recuperación de imágenes.

3.3.3 Observaciones de AIA no supervisada

En la AIA no supervisada se pueden identificar dos subproblemas diferentes: (1) en donde se tiene un imagen a anotar y de su texto asociado se lleva a cabo como una anotación individual, y (2) el enfoque tradicional, en donde para la imagen a anotar no se tiene información de texto asociado y se usa un colección de imágenes de referencia (donde cada imagen cuenta con un texto asociado) para derivar las etiquetas de anotación.

Una de las principales suposiciones de trabajos de AIA no supervisada que realizan la anotación de forma individual [20, 52, 24, 29], es que el texto asociado, el cual para algunos casos sólo es un fragmento pequeño, describe el contenido de la imagen, lo cual no siempre se cumple, ignorando por completo el contenido visual de las imágenes.

En cambio, los trabajos del enfoque tradicional de AIA no supervisada [58, 93, 56, 78, 90, 74], procesan el texto asociado a k imágenes que sirve como información para buscar describir el contenido visual de la imagen a anotar. Sin embargo, dan mayor importancia al contenido visual y llevan a cabo dos pasos en secuencia para realizar la anotación, ignorando las interacciones entre los datos. Además, el paso de minería de texto para extracción de etiquetas depende del paso de CBIR, en donde la recuperación de imágenes no es confiable para todos los casos.

4 Motivación y Justificación

El desarrollo de AIA es de gran utilidad para diversas áreas de investigación que pueden aprovechar la información imagen-etiquetas provista por la anotación para resolver diferentes tareas. Entre las tareas con mayor demanda se encuentra la recuperación de imágenes, debido a que para el incremento de datos en la Web es beneficioso contar con una anotación variada de etiquetas y robusta en las imágenes. Tareas en las que puede influir una correcta anotación de imágenes son el indexado, organización, categorización, reconocimiento, auto-ilustración y exploración de grandes colecciones de imágenes en la Web. Estas tareas se encuentran relacionadas entre sí (trabajos al respecto pueden ser revisados en [23, 6, 69, 47, 66]). Además, el desarrollo de sistemas de AIA no supervisada puede beneficiar a tareas de anotación de lugares geográficos [79, 97], clasificación [10, 18] y robótica [67], que podrían incorporar módulos de anotación como parte integral de los sistemas, por ejemplo procesos de adquisición de imágenes para entrenamiento.

Hemos elegido trabajar AIA bajo un enfoque no supervisado debido a que es posible anotar una mayor diversidad de etiquetas a niveles generales y específicos[6], lo cual es un aspecto deseable ya que permite describir diferentes características en la imagen. Además, usar datos no supervisados conlleva a una mayor escalabilidad en tamaño de base de datos y en número de etiquetas a anotar [15].

Los sistemas actuales en AIA no supervisada usan un módulo de CBIR como principal indicador de relevancia para realizar la anotación, este escenario ocasiona que la anotación de la imagen de entrada recaiga en su apariencia sin importar su contexto [81], lo cual puede ocasionar que se anote incorrectamente debido que no existe actualmente un descriptor visual robusto para una situación general de recuperación.

La mayoría de las estrategias para AIA no supervisada actuales [58, 93, 56, 52, 24, 78, 90, 74, 20] emplean procesos separados para procesar el texto y las imágenes desaprovechando las interacciones entre

[6]Nótese que este aspecto tiene dependencia sobre la calidad y la explotación de la información asociada a la imagen.

las modalidades, más aún no se ha propuesto explotar la información multimodal (texto e imágenes representados en conjunto) para llevar a cabo la tarea de AIA no supervisada. No obstante, la disponibilidad de modalidades incitan al desarrollo de herramientas capaces de aprovechar la diversidad, redundancia y complementariedad de la información [27].

El procesamiento de texto e imágenes como procesos separados en la AIA no supervisada, nos lleva a resaltar la falta de un marco de trabajo que permita introducir nuevos métodos para explotar las relaciones entre el texto y las imágenes, así como datos multimodales. La propuesta de esta investigación consiste en introducir nuevos métodos para mejorar la efectividad de la anotación de imágenes. El escenario en el que nos disponemos a trabajar es aquel en donde dada una imagen y una colección de imágenes de referencia, buscamos asignar etiquetas a la imagen que describan su contenido visual. A diferencia de los métodos actuales nos enfocaremos en la explotación de relaciones de características textuales y visuales, así como el uso de datos multimodales. Para llevar a cabo estos métodos proponemos la definición de AIA no supervisada en analogía con AQE. Un marco de trabajo para analizar la tarea de AIA no supervisada surge de esta analogía y es una contribución de esta investigación. En la definición de este marco de trabajo se analizan e identifican estrategias para subsanar y mejorar la anotación de imágenes. La analogía de AIA no supervisada con AQE nos permite proponer esquemas de anotación de imágenes usando datos multimodales, además de adoptar y definir estrategias empleadas en AQE para evaluar la eficacia en la tarea de anotación de imágenes. Además, un componente novedoso que proponemos analizar y desarrollar bajo el marco de trabajo de anotación de imágenes, es la estimación de la complejidad de la consulta, es decir, proponemos el desarrollo de una métrica para estimar a priori y posteriori la complejidad de llevar a cabo la anotación de una imagen dada usando una determinada colección de imágenes de referencia. El estudio de complejidad es un aspecto deseable que permitirá determinar bajo que paradigma de anotación es posible sacar mayor provecho. Mecanismos para estimar la complejidad de la consulta en AQE han sido estudiados, pero no se han llevado a cabo estudios en AIA no supervisada.

5 Propuesta de Investigación

La propuesta de estudio es investigar, bajo un marco de trabajo definido a partir de la analogía con expansión automática de consultas, a los sistemas de anotación automática de imágenes no supervisados, proponiendo métodos de anotación en paradigmas locales y globales (definidos a partir de la analogía) que incorporen datos multimodales.

La analogía que proponemos nos permite identificar y caracterizar estrategias y métodos para la anotación de imágenes no supervisada, que incluyen el usar información unimodal o multimodal, es decir, en procesamiento textual o visual por separado o en conjunto para el proceso de anotación.

El marco de trabajo de anotación basado en métodos locales y globales es una contribución de esta investigación y abre posibilidades de investigación para analizar el impacto de usar datos multimodales y emplear técnicas de la expansión automática de consultas para resolver la tarea de anotación de imágenes bajo un enfoque no supervisado. En este marco de trabajo se incluye un componente a analizar que no ha sido estudiado anteriormente que es la complejidad de anotar una imagen.

Antes de presentar las preguntas de investigación y los objetivos de la tesis describimos la analogía entre AIA y AQE que tomaremos como base para definir nuestro marco de trabajo para el desarrollo de métodos de anotación no supervisada de imágenes. Nótese que esta analogía es parte de los avances de este propuesta de investigación.

5.1 Analogía entre AIA no supervisada y AQE

Existe una serie de propiedades de correspondencia entre AIA no supervisada y AQE que hemos identificado y que nos permiten ver en analogía a la tarea de AIA no supervisada como una tarea de AQE, estas propiedades se describen a continuación.

En AIA no supervisada una imagen i a anotar puede verse como una consulta q en AQE. A la imagen i se le desean anotar k etiquetas que se busca describan su contenido visual, y en AQE la consulta q se desea expandir con l términos de expansión que buscan reflejar la intención de la consulta que se desea expandir. Nótese que un problema para ambos casos es que k y l son desconocidos, la dificultad de este problema radica en que para describir el contenido visual de una imagen Ψ_i e interpretar la intención de una consulta Ψ_q existe una subjetividad variable. Además, otra característica que puede encontrarse en común es que para determinar k y l es necesario descubrirlos y extraerlos, de un texto asociado para el caso de la AIA no supervisada y de documentos para AQE.

La imagen i está compuesta por descriptores visuales v_i que son como términos t_j que componen a la consulta q. Una colección de imágenes de referencia I puede ser vista como la colección de documentos D. Sin embargo, para la colección de imágenes de referencia de AIA no supervisada se tiene que cada ejemplo en la colección es representado por dos diferentes modalidades, texto e imágenes, en cambio en AQE la colección de documentos D se centra generalmente sólo en texto. Trabajar con dos diferentes modalidades en AIA no supervisada abre posibilidades para tratarlas por separado o en conjunto.

Para llevar a cabo la anotación de una imagen i se utilizan las imágenes de la colección de referencia I, este proceso es como llevar a cabo una expansión, la cual depende de la colección de documentos D. Sin embargo, en estos dos procesos, anotación y expansión, se genera un cambio debido a las modalidades en juego. En AIA no supervisada la anotación se da a una imagen i por medio de texto, $\Phi_i : V \rightarrow T'$, es decir, una función que mapea lo visual a lo textual. En cambio, realizar una expansión en un consulta generalmente es una función que mapea de lo textual a lo textual $\Phi_q : T \rightarrow T'$.

En la Tabla 2 se resumen las características antes mencionadas de la analogía entre AIA no supervisada y AQE.

En AIA no supervisada:	**es como... en AQE:**
Una imagen i.	Una consulta q.
Una imagen esta compuesta por descriptores visuales $i : \{v_1, v_2, \cdots, v_i\}$.	Una consulta esta compuesta por términos $q : \{t_1, t_2, \cdots, t_j\}$.
Desconocido número de conceptos a anotar k.	Desconocido número de términos a expandir l.
Una colección de imágenes de referencia $I : \{i_1a_1, i_2a_2, \cdots, i_na_n\}$.	Una colección de documentos $D : \{d_1, d_2, \cdots, d_m\}$.
Anotación de etiquetas $\Phi_i : V \rightarrow T'$.	Expansión de términos $\Phi_q : T \rightarrow T'$.
La interpretación visual de una imagen Ψ_i.	La intención de una consulta Ψ_q.

Table 2. Analogía entre AIA y AQE

5.2 Arquitectura de AIA no supervisada en paradigmas de AQE

Usando la analogía entre AIA no supervisada y AQE, se proponen dos paradigmas para llevar a cabo la tarea de anotación de imágenes teniendo como base las definiciones de paradigmas de expansión automática de consultas (ver subsección 2.1). La definición de los dos paradigmas de anotación de imágenes busca explotar la información utilizada en el proceso de la anotación, además de agregar la posibilidad a investigar características de interés durante el proceso. Nos centraremos en dos de los más prometedores paradigmas: local y global. Cabe mencionar que los paradigmas de anotación de imágenes que a continuación se presentan, son parte del marco de trabajo que proponemos para introducir estrategias para explotar relaciones entre las características textuales y visuales.

1. **AIA no supervisada local (basada en la consulta)**. Trabajos en este paradigma centran su atención en la información local que puede ser explotada de documentos pseudo-relevantes[7] a la consulta. Los diferentes trabajos revisados de AIA no supervisada pueden ser encuadrados en este enfoque, en donde la anotación depende del procesado del texto asociado a las imágenes similares a la imagen a anotar. Sin embargo, los trabajos actuales no han aprovechado las relaciones entre características textuales, entre características visuales, o en conjunto a través de datos multimodales. Nuestra propuesta para llevar a cabo a la anotación local (puede revisarse en la Figura 8) consiste en:

   - Usar técnicas de análisis de expansión para las características textuales y visuales. El interés de agregar un módulo de expansión es investigar estrategias que puedan mejorar la correspondencia entre los pasos de CBIR y minería de texto. Proponemos investigar módulos de expansión aplicados sólo usando el texto, sólo usando los elementos visuales, ó ambos en conjunto. Cabe mencionar que este módulo es un análisis aplicado *offline* y no interrumpe el proceso de anotación de imágenes a dos pasos.

2. **AIA no supervisada global (basada en el corpus)**. Para este paradigma no se incluye un paso de CBIR, en cambio se lleva a cabo un análisis global del corpus. Proponemos para el paradigma de anotación global:

   - Usar un análisis visual, textual o multimodal de la información de la colección de imágenes de referencia, haciendo uso de representaciones distribucionales. La finalidad de usar el análisis de relaciones y asociaciones de la información es identificar características de las cuales sacar provecho para llevar a cabo la anotación. En la Figura 9 describimos el proceso de anotación usando este paradigma, en donde de inicio es necesario realizar una extracción visual de la imagen a anotar, después se lleva a cabo una búsqueda en la representación de análisis (construida *offline*).

Los paradigmas de anotación anteriores utilizan la información local (de una recuperación de imágenes similares) y global (de una colección de referencia antes de realizar la anotación) para llevar a cabo la anotación. Con estos paradigmas no sólo se busca definir distintos escenarios de AIA no supervisada sino también identificar técnicas que prodrían ser usadas. Así mismo, analizar nichos de investigación que permitan atacar limitaciones de métodos actuales de anotación de imágenes no supervisados, y posibilitar la introducción de técnicas aplicadas en AQE para ser adoptadas en AIA no supervisada. Las limitaciones que se buscan subsanar son definidas en las siguientes preguntas de investigación.

[7] La palabra 'pseudo' hace referencia a que pueden no ser ciertos.

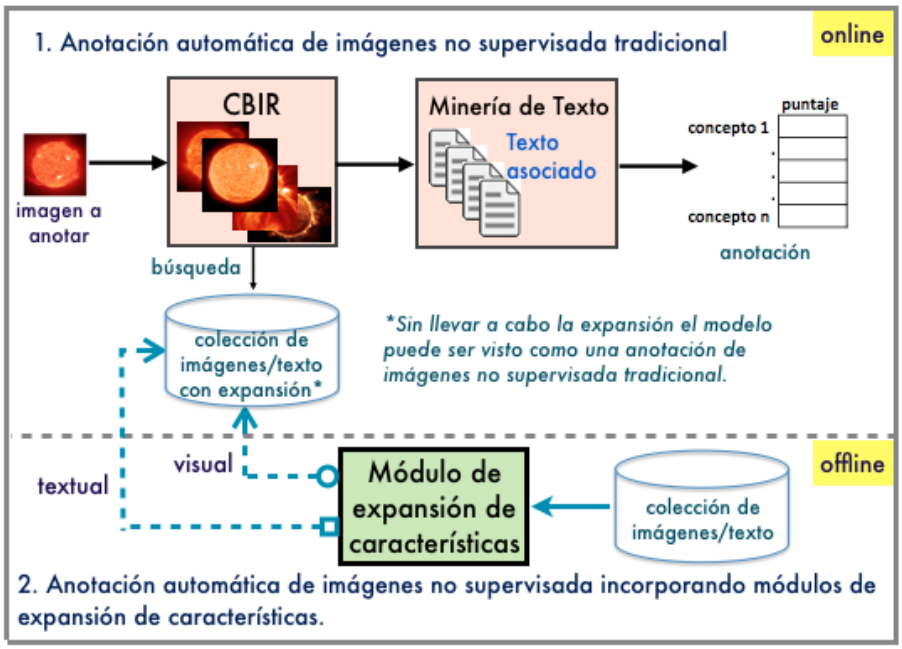

Figure 8. La parte superior de la figura incluye a la tarea de AIA no supervisada como un proceso *online*. En la parte inferior inferior se propone un módulo de expansión de características que ocurre *offline*, la expansión puede ser visual, textual o ambas.

5.3 Preguntas de Investigación

Las preguntas de investigación que perseguimos resolver se encuentran enfocadas a la definición de anotación automática de imágenes no supervisada en analogía a la expansión automática de consultas y son definidas a continuación:

1. **¿Qué ventajas ofrece definir AIA no supervisada en analogía a AQE?**

   Se han definido dos paradigmas de anotación de imágenes a partir de esta analogía: local y global. Con esta pregunta de investigación se investigarán y responderán a el cómo llevar a cabo anotaciones eficaces usando los paradigmas local y global propuestos. Además, la pregunta responderá a qué atributos conviene usar para cada paradigma en evaluación al desempeño y bajo qué condiciones en el proceso de anotación.

2. **¿Qué impacto tiene usar una representación multimodal que capture las relaciones entre atributos textuales y visuales en la anotación de imágenes en comparación a métodos tradiciones que sólo utilizan datos unimodales? y ¿cómo sacar provecho de la representación multimodal en los paradigmas de anotación local y global?**

   Determinar si la información multimodal es de utilidad en comparación con usar datos unimodales, es decir, sólo textual o sólo visual. Evaluar el rendimiento de anotación de imágenes al usar datos multimodales en los paradigmas de anotación local y global.

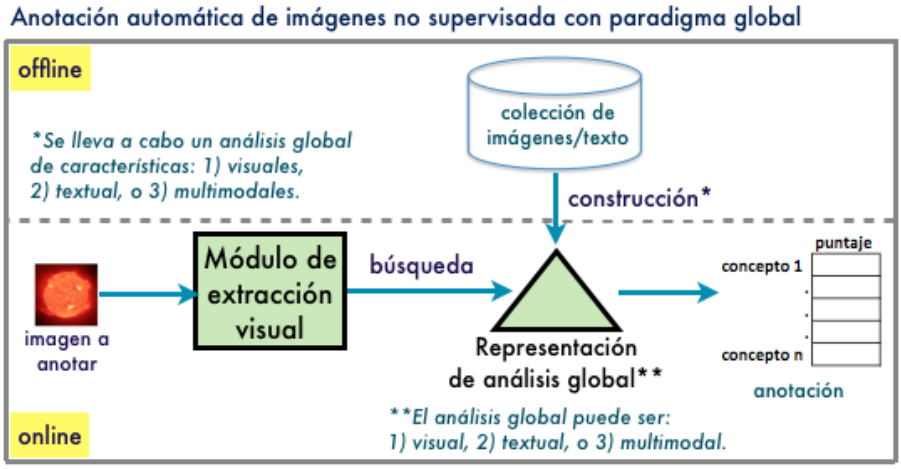

Figure 9. La parte superior de la figura se lleva a cabo un proceso *offline* para llevar a cabo un análisis global de características. En la parte inferior de la figura se presenta el proceso de anotación *online*.

3. **Dada una imagen a anotar y una colección de referencia, ¿bajo qué paradigma local o global conviene realizar la anotación?**

   Dentro de AQE se han definido métricas que permiten identificar la complejidad de la consulta y así determinar el paradigma de expansión adecuado, de manera análoga proponemos definir una métrica para identificar bajo qué paradigma de anotación resulta conveniente realizar la anotación. Comparado con los actuales métodos de AIA no supervisada, el estudio de complejidad es un componente nuevo que no ha sido investigado.

5.4 Objetivos

General

Proponer y desarrollar métodos para anotación no supervisada de imágenes inspirados en paradigmas locales y globales de expansión automática de consultas, que empleen datos multimodales y permitan mejorar la efectividad de métodos de anotación no supervisada del estado del arte.

Específicos

- Definir un marco de trabajo para AIA no supervisada en analogía con AQE.
- Caracterizar el proceso de anotación de imágenes bajo dos paradigmas, local y global, identificando las técnicas de AQE que pueden ser aprovechadas bajo estos paradigmas.
- Proponer métodos de anotación local y global para AIA no supervisada usando la analogía definida con AQE.
- Proponer una representación multimodal entre texto e imágenes que permita capturar las relaciones entre características textuales y visuales, para ser empleada en los paradigmas de anotación de imágenes local y global.

- Proponer métricas para estimar a priori y posteriori la complejidad de la imagen a anotar que permita determinar el paradigma apropiado para llevar a cabo el proceso de anotación de imágenes.

5.5 Metodología

Para alcanzar los objetivos de esta propuesta se presenta la siguiente metodología:

1. Estudio crítico de AQE y AIA no supervisada. Inicialmente y durante el desarrollo de esta metodología se realizará un estudio crítico de AIA no supervisada con la finalidad de identificar los problemas que afectan al desempeño en la anotación. Igualmente, un estudio crítico de AQE para identificar técnicas que puedan ser aprovechadas en la anotación no supervisada y caracterizar el proceso de anotación en analogía con la expansión.

2. Selección de colecciones de imágenes de referencia y fase de preprocesamiento. Primero identificar y seleccionar colecciones de imágenes de referencia que permita evaluar los métodos propuestos. Posteriormente, realizar un procesamiento inicial en las datos textuales y visuales de la colección de imágenes de referencia para usar en etapas posteriores. Identificar posibles fuentes de sesgos en la información al realizar los pasos de procesamiento. Evaluar la posibilidad de crear un base de datos para evaluar los métodos. Algunas de las tareas de preprocesamiento incluyen para texto: extracción del vocabulario, indexado, esquemas de pesado, lematización, etcétera; para imágenes: normalización, extracción de descriptores visuales, indexado, representación en vocabulario visual, entre otras.

3. Desarrollo de método basado en paradigma de expansión local. Primero, realizar una revisión de técnicas aplicadas a paradigmas locales de AQE que puedan ser aprovechadas en la anotación en paradigma local. De acuerdo a las técnicas revisadas proponer una estrategia de anotación local que utilice datos textuales y otra estrategia con datos visuales. Llevar a cabo la experimentación y evaluación para el método de anotación local.

4. Desarrollo de método basado en paradigma de expansión global. Revisar técnicas de análisis global empleadas en paradigmas globales de AQE. Proponer técnicas para análisis de datos tanto textuales como visuales a usar en una anotación global. Definir una estrategia de anotación global y llevar a cabo la experimentación y evaluación del método de anotación global.

5. Experimentación y comparación. Módulo encargado de comparar los métodos de anotación local y global para determinar la efectividad en la anotación al emplear estos paradigmas.

6. Desarrollo de una representación multimodal. Primero, realizar un estudio crítico sobre representaciones multimodales del estado del arte. Propone una representación multimodal entre texto e imágenes que permita capturar las relaciones entre características textuales y visuales. Emplear la representación propuesta bajo los paradigmas de anotación de imágenes local y global. Finalmente llevar a cabo la experimentación y evaluación de los paradigmas usando la representación multimodal.

7. Desarrollo de métrica de visualidad. Definir una métrica para estimar la complejidad de llevar a cabo la anotación de una imagen y colección de imágenes de referencia dadas. La métrica tiene como objetivo caracterizar las propiedades de la colección de referencia y la imagen a anotar, con la finalidad de que dicha caracterización permita estimar la dificultad para llevar a cabo la anotación y determinar que paradigma de anotación local o global se debe seguir.

Para los pasos 3, 4, 5, 6 y 7 se proponen los siguientes experimentos a realizar en la Sección 6, en donde se describen métodos identificados a partir de la revisión del estado del arte en AQE aplicados a cada paso de la metodología.

6 Experimentación

En esta sección se presentan descripciones a los experimentos propuestos. Para cada experimento se define el objetivo, hipótesis, diseño experimental, fuentes potenciales de sesgo y la validación.

6.1 Experimento 1 - metodología pasos 3, 4 y 5: pregunta 1

El experimento 1 esta dividido en tres partes que corresponden a tres pasos de la metodología propuesta.

6.1.1 Parte 1: metodología paso 3

Objetivo: Desarrollar un método de anotación local, incorporando un módulo de expansión de características llevado a cabo mediante un análisis visual o textual. Inicialmente se considerará aplicar un módulo de CBIR para recuperar k imágenes visualmente similares a la imagen a anotar, después un módulo de minería de texto para extraer etiquetas posibles a anotar. Para este experimento se buscará determinar la efectividad de la anotación bajo dos estrategias de expansión:

1. Textual. Realizando un análisis por co-ocurrencias de términos extraídos del texto asociado de las k imágenes, se propone llevar a cabo una expansión del texto asociado buscando mejorar la correspondencia de etiquetas entre las k imágenes recuperadas (ver referente del estado del arte 6.1.1 al final).

2. Visual. Análisis de elementos visuales de las k imágenes, buscando sacar aprovecho de las similitudes a nivel elemento visual, es decir, representando a las imágenes por descriptores visuales para realizar la expansión de elementos visuales.

Hipótesis: Es posible mejorar el rendimiento de anotación local utilizando expansiones de información en comparación a usar un método tradicional sin expansión.
Diseño experimental: El elemento de estudio será el paradigma de anotación local bajo las dos estrategias anteriores. Se considerarán variaciones en el número de k imágenes recuperadas por CBIR, cantidad de texto asociado para medir impacto de expansión, usar diferentes descriptores visuales para realizar la recuperación, el n conceptos a anotar.

6.1.2 Parte 2: metodología paso 4

Objetivo: Desarrollar un método de anotación global que use un análisis de la información visual o textual. Se buscará determinar la efectividad de anotación bajo dos estrategias:

1. Textual. Realizando un análisis textual de la colección de referencia usando diferentes representaciones distribucionales como DOR (*document occurrence representation*) y TCOR (*term cooccurrence representation*) (ver referente del estado del arte 6.1.2 al final).

2. Visual. Análisis de elementos visuales representando a las imágenes por descriptores visuales usando diccionarios visuales (ver referente del estado del arte 6.1.2 al final).

Hipótesis: Es posible mejorar el rendimiento de anotación global realizando un análisis de datos textuales en comparación a análisis en datos visuales.

Diseño experimental: El elemento de estudio será el paradigma de anotación global bajo las dos estrategias anteriores. Se considerarán análisis globales por representaciones textuales y así como visuales de la colección de imágenes de referencia, representación de bolsa de palabras visuales usando diferentes descriptores visuales y el n conceptos a anotar.

6.1.3 Parte 3: metodología paso 5

Objetivo: Determinar la efectividad de anotación para los métodos de anotación propuestos en los paradigmas local y global. Se buscará determinar eficacia de precisión y recuerdo, así como eficiencia bajo diferentes escenarios de prueba:

1. Tamaño de colección de referencia de imágenes.
2. Cantidad de conceptos a anotar.
3. Diversidad en el contenido visual de las imágenes a anotar.
4. Cohesión semántica entre etiquetas anotadas.

Hipótesis: El paradigma de anotación global presenta un mejor rendimiento en anotación que el paradigma de anotación local.

6.2 Experimento 2 - metodología paso 6: pregunta 2

Objetivo: Determinar la efectividad de anotación para los métodos de anotación propuestos en los paradigmas local y global usando una representación multimodal. Se buscará determinar eficacia de precisión y recuerdo, en comparación a paradigmas locales y globales usando solo representaciones tradicionales (unimodales), se realizará una evaluación bajo diferentes escenarios de prueba:

1. Tamaño de colección de referencia de imágenes.
2. Cantidad de conceptos a anotar.
3. Diversidad en el contenido visual de las imágenes a anotar.
4. Cohesión semántica entre etiquetas anotadas.

Para la propuesta de la representación multimodal a evaluar en los paradigmas de anotación de imágenes, y como parte del estudio crítico sobre diferentes representaciones del estado del arte se han identificados referentes teórico de los que tomaran en consideración características para la propuesta (ver referente del estado del arte 6.2 al final).

Hipótesis: El uso de datos multimodales para anotación de imágenes no supervisada tiene mayor eficacia en anotación que usando datos unimodales.

Diseño experimental: Se realizará una comparativa en anotación de los paradigmas de anotación local y global usando datos unimodales y multimodales.

6.3 Experimento 3 - metodología paso 7: pregunta 3

Objetivo: Proponer una métrica para determinar la complejidad de la anotación, dada una imagen y una colección de referencia. Se analizarán propiedades en colección de referencia y en la imagen a anotar a fin de determinar que paradigma de anotación usar (ver referente del estado del arte). Dos perspectivas que son de estudio para este experimento son si es posible determinar la complejidad a priori o a posteriori por medio de los resultados de una anotación inicial.

Hipótesis: Existen características en la colección de referencia que dada una imagen es posible determinar que paradigma de anotación presenta un mejor rendimiento para llevar a cabo la anotación.

Diseño experimental: Se analizará la colección de referencia y la imagen a anotar.

7 Resultados Preliminares

7.1 Estudio crítico de AQE y AIA no supervisada

Parte de la revisión que se ha llevado a cabo sobre AQE y AIA no supervisada se encuentra en los referentes del estado del arte en la definición de los experimentos propuestos para la metodología. Así mismo, avances preliminares se presentan en la definición de la analogía que entre AQE y AIA no supervisada (subsección 5.1), y la definición de la arquitectura de AIA no supervisada en paradigmas de AQE (subsección 5.2).

7.2 Elección de colección de imágenes de referencia y fase de preprocesamiento

Para evaluación del experimento 1, se utilizó el *benchmark* de la subtarea de anotación de conceptos de ImageCLEF13 [93]. Esta colección de imágenes fue creada a partir de 31 millones de consultas usando tres diferentes motores de búsqueda en la Web. Un subconjunto de 250,000 imágenes cada una con un texto asociado fueron seleccionadas para ser usadas como colección de referencia. Otro subconjunto de 1000 imágenes anotadas manualmente con n conceptos de una lista de 107 es usado como conjunto de desarrollo, en donde la lista de los 107 conceptos son utilizados únicamente para evaluación. Las imágenes fueron representadas por una representación vectorial de descriptores visuales: cuatro variantes incluyendo a SIFT, *Color Histogram*, GETLF y GIST, todos ellos mediante la formulación de BoVW *(*bag-of-visual-words).

Para la colección de imágenes de referencia de ImageCLEF se cuenta como texto asociado con las páginas Web en donde fueron encontradas las imágenes. Para el experimento 1, usamos dos tipos de datos asociados: la página Web, considerando todo el texto presente y las palabras clave (*keywords*) que contiene descripciones cortas de las imágenes usadas para la búsqueda.

Para el paso de preprocesamiento la colección de imágenes de referencia fue indexada con los términos presentes en el texto asociado usando TMG 5.0 (Text to Matrix Generator[8]). El indexado se llevo a cabo para los dos recursos texto asociado, después la matrix indexada fue normalizada por imagen. Para reducir y hacer manejable la matriz indexada los *stop words* y los términos de baja frecuencia fueron eliminados.

La medida de evaluación de la anotación es llevada a cabo usando *average precision* (AP). La medida de AP mide el orden inducido por puntajes en la anotación de cada concepto anotado, y es definida como:

$$AP = \frac{1}{|G|} \sum_{g=1}^{|G|} \frac{t}{rank(g)} \tag{1}$$

[8] http://scgroup20.ceid.upatras.gr:8000/tmg/

donde G es el conjunto ordenado de anotaciones *ground truth*, y $rank(g)$ es la posición ordenada de la g-ésima anotación *ground truth*.

7.3 Experimento 1, pregunta 1, parte 1 - metodología paso 3

Método de anotación local simple + expansión textual:
Este experimento forma parte de una serie de experimentos que incorporan estrategias para mejorar el método de anotación de imágenes bajo el paradigma de anotación local que proponemos en analogía con AQE.

(a) Descripción y evaluación:

Para este experimento usamos el proceso de anotación local más simple, el cual consiste en dos módulos (ver Figura 10):

1. Módulo CBIR. Se lleva a cabo una búsqueda en el espacio de características visuales, en donde se recuperan k imágenes con máxima similitud visual de la colección de imágenes de referencia. Se uso para la búsqueda un descriptor visual (por ejemplo SIFT) y L1 como medida de distancia. La definición de la distancia L1 usada fue:

$$L1(x, y) = \frac{1}{D} \sum_{i=1}^{D} |x_i - y_i| \tag{2}$$

   donde x y y son imágenes representadas por un descriptor visual, y D es el tamaño del vocabulario visual. En el experimento se evaluaron diferentes descriptores visuales con el objetivo de comparar el rendimiento de anotación.

2. Módulo de minería de texto. Usa el texto asociado a las k imágenes recuperadas. Para llevar a cabo la anotación se usa una estrategia de mayoría de votos que consiste en anotar con los 10 conceptos de mayor frecuencia en el texto asociado a las k imágenes.

La estrategía de expansión de términos es realizada *offline* y es agregada al proceso de anotación de imágenes (ver Figura 10), el método de expansión es descrito en la subsección (c) de este experimento.

(b) Objetivo:

El objetivo del experimento fue evaluar el impacto de agregar una estrategia de análisis textual que expande el texto asociado, sin cambiar el proceso de anotación local de imágenes. La hipotésis es que los términos que co-ocurren con frecuencia en el texto asociado pueden ser usado como contexto para relacionar a las imágenes. Dadas dos imágenes x y y que comparten términos relacionados la expansión agrega términos de x a y usando un análisis de co-ocurrencias entre los diferentes términos del texto asociado de la colección de imágenes de referencia.

Para evaluar el aprovechamiento de usar la expansión se usaron dos tipos de datos asociados: 1) palabras clave y, 2) páginas Web con diferentes cantidades de información (descritos en la subsección 7.2) y usando un diferente número de k imágenes, también se tomaron en consideración diferentes descriptores visuales para el módulo de CBIR. Nuestros resultados muestran que usando la estrategia de expansión puede ser benéfica para mejorar la eficacia de la anotación bajo diferentes condiciones. Cabe mencionar que la estrategia de expansión no requiere de inferencias complejas o asumir distribuciones en los datos, es una estrategia flexible y puede ser fácilmente combinada al modulo de CBIR sin cambiar el proceso de anotación local. Otros aspectos a destacar de la estrategia son:

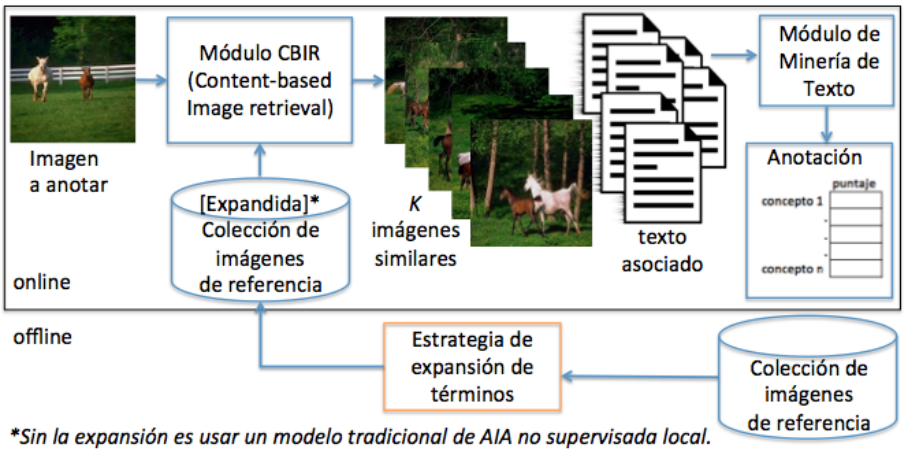

Figure 10. Método de anotación local simple con módulo de expansión de términos.

- Permite definir un indexado de imágenes para consulta tipo *query-by-example* en donde se pueden recuperar imágenes indexadas usando cualquiera de los términos extraídos del texto asociado.
- Provee una anotación relativa a las imágenes de la colección de referencia.
- El modelo de representación tipo recuperación de imágenes, provee de capacidades para analizar relaciones entre imágenes y los términos extraídos del texto asociado.

(c) Estrategia de expansión textual:

La estrategia de expansión aumenta el número de términos en el texto asociado de cada imagen de la colección de referencia. Intuitivamente, dada una imagen a anotar la expansión amplia la información del texto asociado de las k imágenes usadas como referencia para llevar a cabo la anotación, y permite un mayor número de términos para seleccionar como etiquetas de anotación.

La expansión es llevada a cabo por medio de un análisis de co-ocurrencias entre los términos del texto asociado. Para dar una relevancia relativa a los términos extraídos del texto asociado, se estima un peso de las co-ocurrencias entre términos inspirado en una aproximación a la probabilidad condicional:

$$P(y|x) = \frac{O(x,y)}{O(x)} \tag{3}$$

donde $O(x,y)$ expresa el número de veces que los términos x y y occurren juntos, y $O(\cdot)$ expresa el número de ocurrencias de un término. La expansión es llevada a cabo en cada término i extraído del texto asociado con las imágenes indexadas de los términos relacionados (que co-ocurren con i).

Formalmente, se tiene una colección de imágenes de referencia C formada por una conjunto imágenes I con texto asociado T, y se desea llegar a una C' expandiendo T, $C =< I, T > \rightarrow C' =< I, T' >$. Cada imagen en I tiene texto asociado compuesto de n términos $i =< t_1, t_2, \cdots, t_n >$, en donde n es el tamaño del vocabulario de términos. Usando el texto asociado de las imágenes en I se construye una matriz de co-ocurrencias O entre términos, en donde cada término t puede ocurrir m veces con los n términos del

vocabulario, para el término 1 se tiene $t_1 =< t_1^1, t_1^2, \ldots, t_1^n >$. El peso para cada término es calculado usando la ecuación 3. La expansión para cada término de una imagen dada es como sigue:

$$i'_t = \sum_{j=1}^{m} t_j \cdot i_{t_j} \tag{4}$$

donde m es el número de términos que co-ocurren con el término a expandir en O, y i_{t_j} expresa el valor de frecuencia del término j en la imagen.

Para llevar a cabo la expansión sólo consideramos los términos con mayor puntaje de co-ocurrencia. Se filtraron para cada términos aquellos términos relacionados con mayor puntaje de co-ocurrencia después de aplicar la ecuación 3, es decir, se llevo a cabo la expansión usando la ecuación 4 con los términos que pasan un determinado umbral. Se tomaron en consideración varios umbrales, eligiendo finalmente un valor dinámico que depende de la media y la desviación estándar sobre los valores de peso para cada término en la co-ocurrencia. Encontramos que establecer un umbral fijo óptimo era difícil, la razón es que los conceptos asignados a diferentes imágenes contribuyen en diferentes rangos.

(d) Resultados:

Dividimos las pruebas en dos partes. Nuestra primera parte fue evaluar el proceso de anotación local con los dos tipos de texto asociado para establecer los resultados de partida para la calidad de información de cada texto asociado. Posteriormente agregamos nuestra extrategia de expansión de términos y evaluamos el impacto.

Los resultados son mostrados en la Figura 11. Hemos considerado diferentes descriptores visuales para comparar su rendimiento en anotación. En la Figura 11, podemos observar claramente un mejor rendimiento en anotación cuando usamos la página Web completa como texto asociado para los siete descriptores visuales evaluados (ver Figure 11 (a)). Creemos que la página Web contiene mayor información para describir el contenido visual de las imágenes, a diferencia de las *keywords* como texto asociado que sólo contiene pocos términos.

Para los diferentes descriptores visuales podemos observar un ventaja en la familia de descriptores visuales SIFT (ver Figure 11 (b)). Hemos considerado la familia de descriptores visuales SIFT para los siguientes resultados, en particular OPPONENT-SIFT que ha mostrado mayor eficacia.

En la siguiente parte de pruebas hemos agregado nuestra estrategia de expansión textual (descrita anteriormente en la subsección (c)), y hemos conservado el mismo proceso de anotación. Los resultados se muestran en la Figura 12, en donde comparamos el rendimiento de anotación sin expansión, usando nuestra estrategia de expansión, y una expansión total tomando en cuenta todos los términos que co-ocurren considerando *keywords* y páginas Web completas.

Sin usar expansión, podemos observar que usando *keywords* (Figura 12(b)) el rendimiento en la anotación tiende a incrementar con mayor lentitud después de considerar $k \approx 256$ imágenes, consideramos que este comportamiento es debido a la saturación de información. En el rendimiento de anotación usando páginas Web se puede ver que tiene un lapso de convergencia después de considerar $k \approx 256$ imágenes (Figura 12(a)), creemos que la convergencia se debe a la saturación de información.

En cambio, usando una expansión completa se puede observar que usando *keywords* (Figura 12(b)) ha mejorado con la estrategia de expansión, mientras que la anotación usando páginas Web (Figura 12(a)) ha bajado su eficacia conforme se incrementa el número de k imágenes. El comportamiento de anotación usando páginas Web puede ser visto como normal, puede ser intuitivo pensar que información de las páginas Web esta lo suficientemente enriquecida para necesitar expansión.

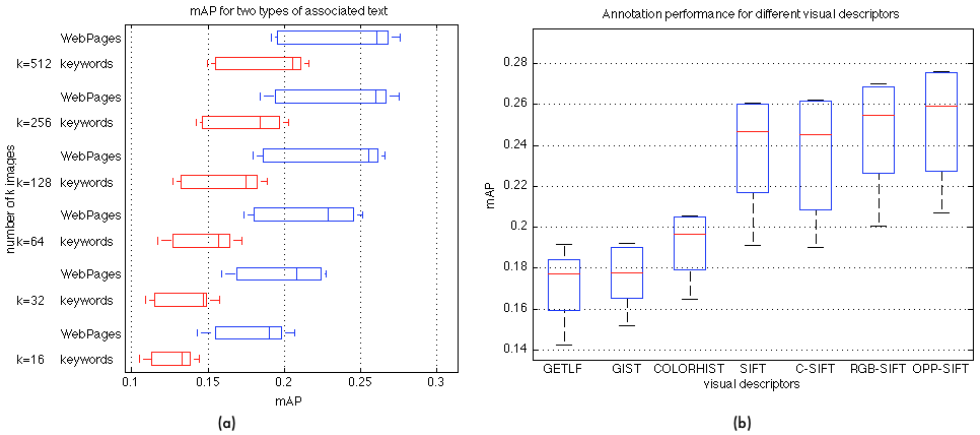

Figure 11. Rendimiento de anotación expresada en mAP (*mean Average Precision*): (a) usando *keywords* y la página Web completa como texto asociado y usando 7 descriptores visuales, y (b) los 7 diferentes descriptores visuales tomando en cuenta el rendimiento tanto de *keywords* como de páginas Web.

Finalmente, usando nuestro método de expansión, podemos observar que al usar un menor número de término en la expansión en páginas Web (Figura 12(a)), en este caso un umbral con la media + desviación estándar, hemos mejorado la anotación usado pocas imágenes, incluso alcanzado el máximo rendimiento usando $k = 512$. No obstante, podemos ver de la Figura 12 (a) que cuando consideramos grandes valores para k (e.g., mayores a 128) el rendimiento de anotación usando la estrategia de expansión decae. Creemos que el decremento de eficacia es debido a que la información es ruidosa para minar los conceptos a anotar. Por otro lado, se puede observar un rendimiento de anotación similar al usar una expansión completa y nuestra estrategia de expansión para *keywords* Figura 12 (b). Nuestro método de expansión empieza con un menor rendimiento que la expansión total pero después de $k = 128$ ambas estrategias incrementan a la par, creemos que el rendimiento de anotación en este caso es similar debido que se tienen muy pocos términos en el texto asociado de *keywords*.

Se ha mencionado que el texto asociado de *keywords* contiene pocos términos, a continuación en la Figura 13 mostramos resultados del método de expansión propuesto. Bajo la leyenda *original* se encuentran los términos extraídos, mientras que bajo la leyenda *expansion* se muestran los términos resultantes. Para el caso de la expansión se muestran top 10 de términos con mayor puntaje, sin embargo, en algunos casos no se logran conseguir 10 términos (por ejemplo con las imágenes que tienen como texto asociado 'mallard', 'siberian' y 'adulthood'). Es interesante observar que términos en la expansión guardan relación con términos antes de la expansión, habilitando la posibilidad de nuevas opciones de etiqueta a anotar. Por ejemplo, usando {'chortle', 'effectively', 'exposes'} tercer imagen del primer renglón, de la expansión se obtiene 'comedian' y 'guffaw'. Otra observación de interés es que algunos términos pueden ser ambiguos como 'siberian' (segunda imagen primer rengón) y expanden diferentes conceptos, como 'cat' y 'purring', que no están relacionados con el contenido visual de la imagen.

Por otro lado, podemos evaluar de manera cualitativa la expansión listando las imágenes antes y después de aplicar la expansión para un término o concepto dado. En la Figura 14 presentamos top 10 de imágenes relacionadas con el concepto 'cloud', en la parte superior de la línea roja en la figura se muestran

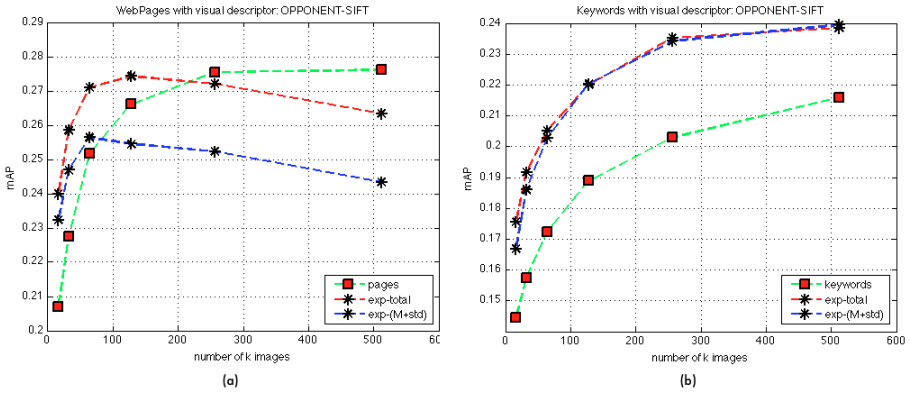

Figure 12. Rendimiento de anotación en *mean Average Precision* (mAP): (a) usando páginas Web como texto asociado vs. versiones expandidas usando el descriptor visual OPPONENT-SIFT, y (b) usando *keywords* como texto asociado vs. versiones expandidas usando el descriptor visual OPPONENT-SIFT.

las 10 imágenes correspondientes antes de la expansión en orden de izquierda a derecha, siguiendo un ordenamiento similar en la parte inferior de la línea roja se muestran las 10 imágenes correspondientes aplicando la expansión. Podemos observar que antes de llevar a cabo la expansión hay imágenes que no están relacionadas con el concepto pero después de llevar a cabo la expansión esas imágenes han sido desplazadas a niveles inferiores, en cambio, la expansión ha favorecido a agregar imágenes en las que se pueden diferenciar tres tópicos: 1) 'cloud' visto como definición de meteorología (imágenes 2, 4, 5, 6, 7 y 10), 2) 'cloud' relacionada a computación (imágenes 1 y 8), y 3) 'cloud' relacionada a un video juego (imágenes 3 y 9).

Después de observar el primer listado con 'cloud', es importante señalar que la expansión lleva a cabo dos funciones: (1) agregar nuevas imágenes en las principales tópicos relacionados al concepto, y (2) asignar una relevancia relativa determinada por las asociaciones entre términos para cada tópicos relacionado al concepto.

En la Figuras 15 y 16 se muestran dos ejemplos más del antes y después de llevar a cabo la expansión. Usando el concepto 'traffic' (Figura 15), presentamos el top 10 de imágenes relacionadas con el concepto, en la parte superior de la línea roja se muestran las 10 imágenes correspondientes antes de la expansión en orden de izquierda a derecha, siguiendo un ordenamiento similar en la parte inferior de la línea roja se muestra las 10 imágenes correspondientes aplicando la expansión. Podemos observar un claro incremento de imágenes con relación al concepto, y se distinguen dos tópicos: 1) 'traffic' con referencia a transportación, y 2) 'traffic' relacionado a Internet.

En la Figura 16 usando el concepto 'mountain' presentamos el top 10 de imágenes relacionadas con el concepto, en la parte superior de la línea roja se muestran las 10 imágenes correspondientes antes de la expansión en orden de izquierda a derecha, siguiendo un ordenamiento similar en la parte inferior de la línea roja se muestra las 10 imágenes correspondientes aplicando la expansión. Podemos observar que se ha llevado un incremento en imágenes que muestran montañas sobre los demás tópicos que puedan estar relacionados con el concepto favoreciendo a la anotación del concepto.

(f) Conclusiones del experimento:

Figure 13. Anotación antes y después de aplicar la expansión.

Los resultados preliminares obtenidos en esta primera parte del experimento 1 resultan alentadores, con sólo agregar una simple estrategia textual que analiza el contexto de las imágenes ha sido posible mejorar el desempeño de anotación de un paradigma local. Los siguientes pasos para completar el experimento 1 son incluir una estrategia de expansión que use elementos visuales, incluir una evaluación comparativa entre las dos modalidades describiendo bajo que condiciones resulta benéfico utilizar expansiones usando características textuales o visuales. Finalmente evaluar el paradigma de anotación de imágenes local incluyendo un proceso iterativo de pseudo retroalimentación de relevancia.

Derivado de este experimento, parte 1, se ha sometido un artículo titulado: *Evaluating Term-Expansion for Unsupervised Image Annotation* en MICAI'14 (*Mexican International Conference on Artificial Intelligence*).

(g) Tablas de resultados del experimento 1:

A continuación se muestran tablas de resultados del experimento 1, de las cuales se obtuvieron las gráficas presentadas.

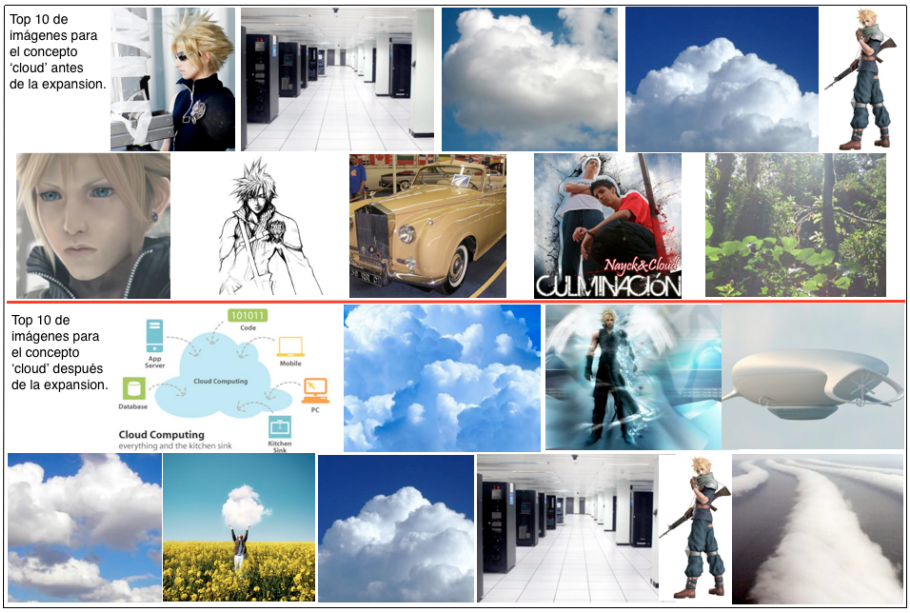

Figure 14. Top 10 imágenes relacionadas con el concepto 'cloud' antes y después de la expansión.

Figure 15. Top 10 imágenes relacionadas con el concepto 'traffic' antes y después de la expansión.

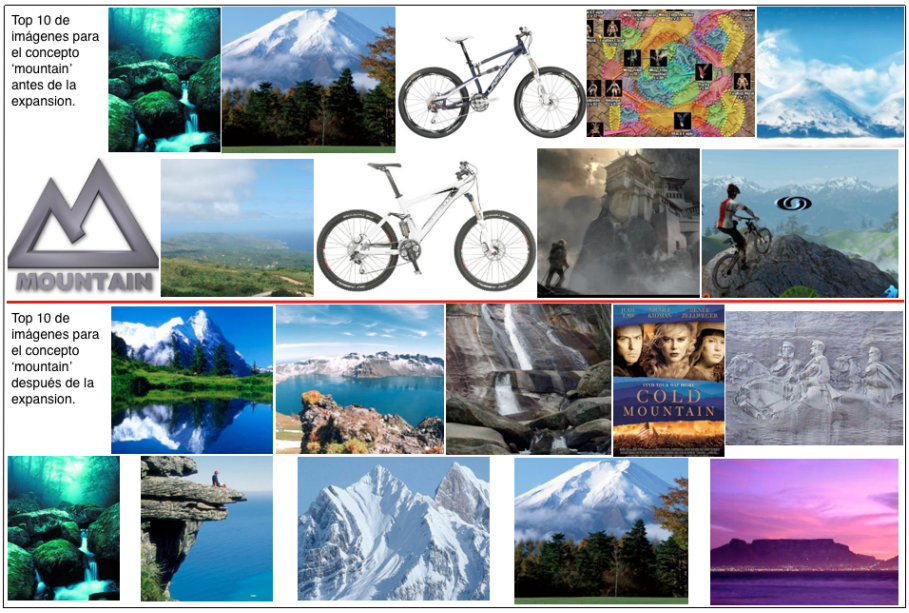

Figure 16. Top 10 imágenes relacionadas con el concepto 'traffic' antes y después de la expansión.

Descriptor visual	**k**= 16	32	64	128	256	512
OPP-SIFT[1]	0.1446	0.1575	0.1722	0.1889	0.203	0.216
RGB-SIFT[1]	0.1397	0.1494	0.1652	0.1832	0.1982	0.2115
C-SIFT[1]	0.1332	0.1468	0.1572	0.1748	0.1844	0.2079
SIFT[1]	0.1343	0.1468	0.1626	0.1792	0.1922	0.2055
ColorHist[1]	0.1155	0.1227	0.1336	0.1442	0.1568	0.1604
GIST[1]	0.1122	0.1123	0.125	0.1281	0.143	0.1529
GETLF[1]	0.1053	0.1093	0.1171	0.1272	0.1425	0.1494
OPP-SIFT[2]	0.2069	0.2274	0.2517	0.2661	0.2755	0.2761
RGB-SIFT[2]	0.2006	0.2266	0.2481	0.2617	0.2686	0.2702
C-SIFT[2]	0.1903	0.2084	0.2288	0.2613	0.2618	0.262
SIFT[2]	0.191	0.2171	0.2384	0.2552	0.2602	0.2607
ColorHist[2]	0.165	0.1795	0.1921	0.2014	0.2053	0.2055
GIST[2]	0.152	0.1654	0.1759	0.1794	0.1902	0.1924
GETLF[2]	0.1427	0.1592	0.1734	0.1813	0.1841	0.1916

Table 3. Rendimiento de mAP para diferentes descriptores visuales usando diferente número de k imágenes en dos tipos de texto asociado: 1) *keywords* y, 2) páginas Web.

Visual Descriptor	**k**= 16	32	64	128	256	512
OPP-SIFT[1]	0.2069	0.2274	0.2517	0.2661	**0.2755**	**0.2761**
RGB-SIFT[1]	0.2006	0.2266	0.2481	0.2617	**0.2686**	**0.2702**
C-SIFT[1]	0.1903	0.2084	0.2288	**0.2613**	**0.2618**	**0.262**
SIFT[1]	0.191	0.2171	0.2384	0.2552	**0.2602**	**0.2607**
ColorHist[1]	0.165	0.1795	0.1921	0.2014	**0.2053**	**0.2055**
GIST[1]	0.152	0.1654	0.1759	0.1794	**0.1902**	**0.1924**
GETLF[1]	0.1427	0.1592	0.1734	0.1813	0.1841	**0.1916**
OPP-SIFT[2]	0.2323	0.2469	0.2564	0.2545	0.2523	0.2433
RGB-SIFT[2]	0.2178	0.2373	0.2469	0.2448	0.2475	0.2418
C-SIFT[2]	0.2063	0.2174	0.2251	0.2355	0.2318	0.222
SIFT[2]	0.2142	0.2305	0.2418	0.2415	0.2376	0.2299
ColorHist[2]	0.1747	0.1892	0.1936	0.1942	0.1903	0.1884
GIST[2]	0.1655	0.1742	0.1725	0.1732	0.1727	0.1681
GETLF[2]	0.1551	0.1664	0.1788	0.1814	0.1767	0.1753
OPP-SIFT[3]	**0.2399**	**0.2585**	**0.271**	**0.2743**	0.272	0.2633
RGB-SIFT[3]	**0.228**	**0.2501**	**0.2569**	**0.2635**	0.2651	0.2643
C-SIFT[3]	**0.2181**	**0.2267**	**0.2428**	0.2502	0.2499	0.2413
SIFT[3]	**0.2241**	**0.2458**	**0.2588**	**0.261**	0.2575	0.2505
ColorHist[3]	**0.1823**	**0.1971**	**0.2015**	**0.2021**	0.1973	0.1933
GIST[3]	**0.1723**	**0.1813**	**0.1875**	**0.1837**	0.1848	0.1818
GETLF[3]	**0.1611**	**0.174**	**0.185**	**0.1863**	**0.1882**	0.1866

Table 4. Rendimiento de mAP para diferentes descriptores visuales usando diferente número de k imágenes en tres configuraciones de texto asociado: 1) páginas Web, 2) expansión en páginas Web y 3) expansión en páginas Web usando como umbral mean + std, el mejor rendimiento de mAP es mostrado en negritas.

8 Conclusiones

La anotación automática de imágenes es un tema de gran interés científico que se encuentra relacionado con diferentes tareas que involucran el uso de imágenes. Diversas investigaciones han surgido bajo diferentes enfoques con el objetivo de proveer sistemas para realizar la anotación. Investigaciones recientes en la AIA no supervisada no han aprovechado las relaciones entre las diferentes modalidades, visuales y textuales, para llevar a cabo la anotación, y no han considerado el uso de éstos datos en conjunto.

En esta investigación proponemos abordar la anotación no supervisada de imágenes en analogía a expansión automática de consultas. Esta manera de abordar la tarea de AIA no supervisada no ha sido investigada anteriormente. Bajo este nuevo marco de trabajo nos proponemos investigar el impacto de utilizar datos multimodales en el proceso de anotación bajo dos diferentes paradigmas. Cabe mencionar que la analogía y el marco de trabajo propuestos proveen una cama de pruebas para investigar diferentes factores que intervienen durante el proceso de anotación no supervisada que no han sido considerados. Además, de que proponemos un nuevo componente de estudio: un métrica para estimar la complejidad para llevar a cabo la anotación de una imagen.

Las principales contribuciones esperadas con este tema de investigación son aportar conocimiento en la anotación de imágenes no supervisada a través de: (i) caracterizar el proceso de anotación mediante un nuevo marco de trabajo en analogía con expansión automática de consultas para identificar factores que contribuyen la brecha semántica, (ii) proponer nuevos métodos de anotación capaces de sacar provecho de datos multimodales durante el proceso, (iii) un análisis comparativo entre dos diferentes paradigmas y bajo qué circunstancias conviene utilizarlos, y (iv) una métrica para determinar la complejidad de llevar a cabo la anotación de una imagen utilizando una determinada colección de referencia.

8.1 Plan de publicaciones

De la investigación que proponemos para AIA no supervisada se esperan las siguientes publicaciones:

- Las expansiones para métodos de anotación local y global, pueden generar al menos dos publicaciones de conferencia. El resultado que buscamos en particular proponiendo estos dos paradigmas de anotación es mejorar la eficacia en la asignación de conceptos.
- La analogía entre AIA no supervisada y AQE contribuye a la definición de un marco de trabajo que contiene a los dos paradigmas de anotación que proponemos, incluyendo técnicas de AQE que puede ser adoptadas como estrategias para explotar relaciones entre datos textuales y visuales. De este tema se derivaría una publicación de revista.
- El proceso de AIA no supervisada es llevada a cabo por dos pasos principales que ignoran la información en conjunto entre texto e imágenes. Se propone un representación multimodal para sacar provecho de la información conjunta entre el texto y las imágenes. De este tema se derivaría al menos una publicación.
- La anotación de imágenes no es una tarea trivial por lo que es necesario desarrollar herramientas que ayuden a identificar características problemáticas para sobrellevar el proceso de anotación de una manera eficaz. Por lo tanto, se propone una métrica para estimar la complejidad de la imagen y sacar provecho en la utilización de un paradigma de anotación. De este tema se derivaría una publicación.

A continuación algunas de los congresos y revistas consideradas para publicaciones de los productos obtenidos para la investigación propuesta en este reporte.

1. Artículos de congreso.

    - Conferencia: MICAI (*Mexican International Conference on Artificial Intelligence*). Cuenta con tópicos en diversas áreas en Inteligencia Artificial.
    - Conferencia: CVPR (*Conference on Computer Vision and Pattern Recognition*). La anotación de imágenes es un área de investigación fuertemente activa en esta conferencia.
    - Conferencia: ICPR (*International Conference on Pattern Recognition*). Esta conferencia cuenta con un *track* en análisis multimedia, indexado y recuperación.
    - Conferencia: ACM Multimedia. Conferencia especializada en el tema multimedia.
    - Conferencia: ACM ICMR (*International Conference on Multimedia Retrieval*). Conferencia sobre recuperación usando datos multimedia.
    - Conferencia: ICCV (*International Conference on Computer Vision*).

2. Artículos de revista.

    - Revista: *Multimedia Systems*. Revista especializada en sistemas multimedia.
    - Revista: *Multimedia Tools and Applications*. Revista enfocada en herramientas y aplicaciones multimedia.
    - Revista: *Computer Vision and Image Understanding*. Revista enfocada en análisis de imágenes.
    - Revista: *International Journal of Computer Vision*. Tiene Visión por Computadora como tópico de investigación.

8.2 Cronograma de actividades

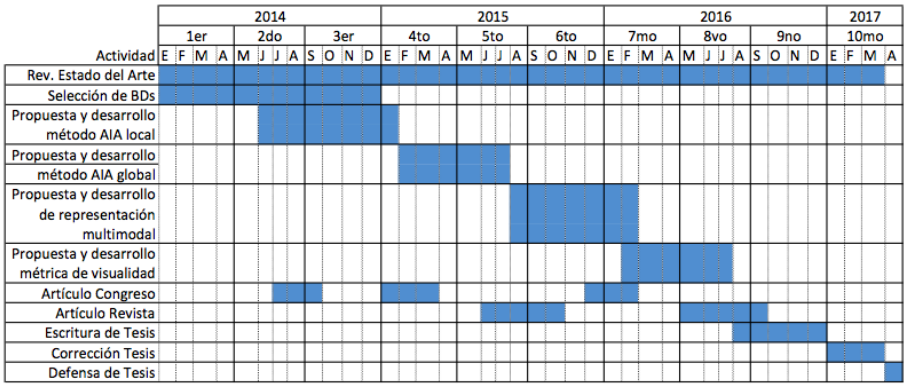

Figure 17. Cronograma de actividades

References

[1] Atrey P. K., Hossain M. A., El Saddik A. and Kankanhalli M. S. (2010). Multimodal fusion for multimedia analysis: a survey. In: *Multimedia Systems, Springer-Verlag*, Vol. 16, Issue 6, 345-379.

[2] Ballan L., Bertini M., Uricchio T. and Del Bimbo A. (2013) Social Media Annotation. In: *11th International Workshop on Content-Based Multimedia Indexing (CBMI)*, 229-235.

[3] Ballan L., Uricchio T., Seidenari L. and Del Bimbo A. (2014) A Cross-media Model for Automatic Image Annotation. In: *Proceedings of International Conference on Multimedia Retrieval (ICMR)*, 73-81.

[4] Bao B. K., Li T. and Yan S. (2012). Hidden-Concept Driven Multilabel Image Annotation and Label Ranking. In: *IEEE Transactions on Multimedia*, Vol. 14, Issue 1, 199-210.

[5] Barnard K., Duygulu P., Forsyth D., de Freitas N., Blei D. and Jordan M. (2003) Matching Words and Pictures. In: *Journal of Machine Learning Research*, Vol. 3, 1107-1135.

[6] Barnard K. and Forsyth D. (2000) Learning the Semantic of Words and Pictures. In: *International Conference on Computer Vision*, Vol. 2, 408-415.

[7] Benavent X., Castellanos A., de Ves E., Hernández-Aranda D., Granados R., Garcia-Serrano A. (2013) A multimedia IR-based system for the Photo Annotation Task at ImageCLEF2013. In: *CLEF 2013 Evaluation Labs and Workshop*, Online Workings Notes. Valencia, Spain.

[8] Bengio Y., Courville A., and Vincent P. (2013) Representation Learning: A Review ad New Perspectives. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 35, No. 8, 1798-1828.

[9] Blei D.M. and Jordan M.I. (2003) Modeling annotated data. In: *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, 127-134.

[10] Boiman O., Shechtman E. and Irani M. (2008) In Defense of Nearest-Neighbor Based Image Classification. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1-8.

[11] le Borgne H., Popescu A. Znaidia A. (2013). CEA LIST@imageCLEF 2013: Scalable Concept Image Annotation. In: *CLEF 2013 Evaluation Labs and Workshop*, Online Workings Notes. Valencia, Spain.

[12] Bruni E., Tran N.K. and Baroni (2014) Multimodal Distributional Semantics. In: *Journal of Artificial Intelligence Research (JAIR)*, Vol. 49, 1-47.

[13] Caicedo J. C. , BenAbdallah J., González F. and Nasraoui O. (2012) Multimodal Representation, Indexing, Automated Annotation and Retrieval of Image Collection via Non-negative Matrix Factorization. In: *Neurocomputing*, Vol. 76, Issue 1, 50-60.

[14] Caicedo J. C. and González F. A. (2012). Online Matrix Factorization for Multimodal Image Retrieval. In: *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, LNCS Vol. 7441, 340-347.

[15] Carneiro G., Chan A., Moreno P. and Vasconcelos N. (2007) Supervised Learning of Semantic Classes for Image Annotation and Retrieval. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 29, Issue 3, 394-410.

[16] Carpineto C. and Romano G. (2012) A Survey of Automatic Query Expansion in Information Retrieval. In: *ACM Computing Surveys (CSUR)*, Vol. 44, Issue 1, Article 1.

[17] Chandrika P. and Jawahar C. (2010) Multi Modal Semantic Indexing for Image Retrieval. In: *Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR)*, 342-349.

[18] Chen M. and Hauptmann (2004) Multi-modal Classification in Digital News Libraries. In: *Conference on Digital Libraries. Proceedings of the 2004 Joint ACM/IEEE*, 212-213.

[19] Chen G., Song Y., Wang F. and Zhang C. (2008) Semi-supervised Multi-label Learning by Solving a Sylvester Equation. In: *Proceedings of the 2008 SIAM International Conference on Data Mining*, 410-419.

[20] Choi D. and Pankoo K. (2012) Automatic image Annotation Using Semantic Text Analysis. In: *Multidisciplinary Research and Practice for Information Systems*, LNCS Vol. 7465, 479-487.

[21] Cronen-Townsend S. and Croft W.B. (2002) Quantifying Query Ambiguity. In: *Proceedings of the second international conference on Human Language Technology Research*, 104-109.

[22] Cusano C., Ciocca G. and Schettini R. (2003) Image annotation using SVM. In: *Proceedings of SPIE 5304, Internet Imaging V*. Vol. 5304, 330-338.

[23] Datta R., Joshi D., Li J. and Wang J.Z. (2008) Image Retrieval: Ideas, Influences, and Trends of the New Age. In: *ACM Computing Surveys (CSUR)*, Vol. 40, Issue 2, Article 5.

[24] Deschacht K. and Moens M.F. (2007) Text Analysis for Automatic Image Annotation. In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, LNCS Vol. 7465, 479-487.

[25] Duygulu P., Barnard K., de Freitas N., and Forsyth D. (2002) Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: *Proceedings of the 7th European Conference on Computer Vision (ECCV)*, Part IV, 97-112.

[26] Escalante H.J. (2010). Cohesión Semántica para la Anotación y Recuperación de Imágenes. *PhD Thesis*, Instituto Nacional de Astrofísica, Óptica y Electrónica.

[27] Escalante H.J., González J.A., Hernández C.A. López A., Montes M., Morales E., Sucar L.E. and Villaseñor L. (2009). Annotation-Based Expansion and Late Fusion of Mixed Methods for Multimedia Image Retrieval. In: *Evaluating Systems for Multilingual and Multimodal Information Access*, LNCS, Vol. 5706, 669-676.

[28] Escalante H.J., Montes M. and Sucar E. (2012) Multimodal Indexing based on Semantic Cohesion for Image Retrieval. In: *Information Retrieval*,Vol. 15, Issue 1, 1-32.

[29] Feng Y. and Lapata M. (2008) Automatic Image Annotation Using Auxiliary Text Information. In: *Proceedings of Annual Meeting of the Association of Computational Linguistics*, 272-280.

[30] Gavves E., Snoek C.G.M. and Smeulders A.W.M. (2012) Visual synonyms for landmark image retrieval. In: *Computer Vision and Image Understanding*, Vol. 116, Issue 2, 238-249.

[31] van Gemert J.C., Veenman C.J., Smeulders A.W.M. and Geusebroek J.M. (2010) Visual Word Ambiguity. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 32, Issue 7, 1271-1283.

[32] Grana C., Serra G., Manfredi M., Cucchiara R., Martoglia R. and Mandreoli F. (2013) UNIMORE at ImageCLEF 2013: Scalable Concept Image Annotation. In: *CLEF 2013 Evaluation Labs and Workshop*, Online Workings Notes. Valencia, Spain.

[33] Grauman K. and Leibe B. (2011) Visual Object Recognition. In: *Synthesis Lectures on Artificial Intelligence and Machine Learning*, Morgan & Clayton Publishers.

[34] Guillaumin M., Mensink T., Verbeek J. and Schmid C. (2009) TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation. In: *IEEE 12th International Conference on Computer Vision*, 309-316.

[35] Guillaumin M., Verbeek J., Schmid C. (2010). Multimodal semi-supervised learning for image classification. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*, 902-909.

[36] Hanbury A. (2008) A Survey of Methods for Image Annotation. In: *Journal of Visual Languages & Computing*, Vol. 19, Issue 5, 617-627.

[37] Hare J. S., Lewis P. H., Enser P. G. B. and Sandom C. (2006) Mind the Gap: Another look at the problem of the semantic gap in image retrieval. In: *Proceedings of Multimedia Content Analysis, Management, and Retrieval (SPIE)*, 607309-1.

[38] Hentschel C., Stober S., Nürnberger A. and Detyniecki M. (2008) Automatic Image Annotation Using a Visual Dictionary Based on Reliable Image Segmentation. In: *Adaptive Multimedia Retrieval: Retrieval, User, and Semantics*, LNCS Vol. 4918, 45-56.

[39] Hidaka M., Gunji N. and Harada T. (2013) MIL at ImageCLEF 2013: Scalable System for Image Annotation. In: *CLEF 2013 Evaluation Labs and Workshop*, Online Workings Notes. Valencia, Spain.

[40] Inkpen D. Z. and Hirst G. (2002) Acquiring Collocations for Lexical Choice between Near-Synonyms. In: *Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition (ULA)*, Vol. 9, 67-76.

[41] Irie G., Liu D., Li Z., Chang S. F. (2013). A Bayesian Approach to Multimodal Visual Dictionary Learning. In: *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 329-336.

[42] Jamil N. and Sa'adan S. A. (2009) Automatic image Annotation Using Color K-Means Clustering. In: *Visual Informatics: Bridging Research and Practice*, LNCS Vol. 5857, 645-652.

[43] Jeon J., Lavrenko V. and Manmatha R. (2003). Automatic Image Annotation and Retrieval using Cross-Media Relevance Models. In: *Proceedings of the 26th annual international ACM Special Interest Group on Information Retrieval (SIGIR)*, 119-126.

[44] Jeon, J. and Manmatha, R. (2004). Using Maximum Entropy for Automatic Image Annotation. In: *International Conference on Content based Image and Video Retrieval (CIVR)*, LNCS Vol. 3115, 24-32.

[45] Jeong J.W. and Lee D.H. (2014) Automatic image annotation using affective vocabularies: Attribute-based learning approach. In: *Journal of Information Science*, Vol. 40, No. 4, 426-445.

[46] Jia Y., Salzmann M. and Darrell T. (2011) Learning Cross-modality Similarity for Multinomial Data. In: *Proceedings of the International Conference on Computer Vision (ICCV)*, 2407-2414.

[47] Joshi D., Wang J. and Li W. (2006). The Story Picturing Engine - A System for Automatic Text Illustration. In: *ACM Transactions on Multimedia Computing, Communications and Applications (TOMM)*, Vol. 2, Issue 1, 68-89.

[48] Krapac J., Allan M., Verbeek J. and Jurie F. (2010) Improving Web Image Search using Query-Relative Classifiers. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1094-1101.

[49] Kherfi M.L. and Ziou D. (2004) Image Retrieval From the World Wide Web: Issues, Techniques, and Systems. In: *ACM Computing Surveys (CSUR)*, Vol. 36, Issue 1, 35-67.

[50] Lavelli A., Sebastiani F. and Zanoli R. (2004) Distributional Term Representations: An Experimental Comparison. In: *Proceedings of the thirteenth ACM International Conference on Information and Knowledge Management (CIKM)*, 615-624.

[51] Lavrenko V., Manmatha R. and Jeon J. (2003) A model for learning the semantics of pictures. In: *Proceedings of the 16th Conference on Advances in Neural Information Processing Systems (NIPS)*, 553-560.

[52] Leong C.W., Mihalcea R. and Hassan S. (2010) Text Mining for Automatic Image Tagging. In: *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, 647-655.

[53] Li B., Duan L.Y., Chen Y., Ji R. and Gao W. (2012) Predicting the Effectiveness of Queries for Visual Search. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2361-2364.

[54] Li X., Gavves E., Snoek C.G.M., Worring M. and Smeulders A. W.M. (2011) Personalizing Automated Image Annotation using Cross-Entropy. In: *Proceedings of the 19th ACM international conference on Multimedia (MM)*, 233-242.

[55] Li X., Liao S., Liu B., Yang G., Jin Q., Xu J., Du X. (2013) Renmin University of China at ImageCLEF 2013 Scalable Image Annotation. In: *CLEF 2013 Evaluation Labs and Workshop*, Online Workings Notes. Valencia, Spain.

[56] Li X., Snoek C.G.M. and Worring M. (2009) Learning Social Tag Relevance by Neighbor Voting. In: *IEEE Transactional on Multimedia*, Vol. 11, Issue 7, 1310-1322.

[57] Liu Y., Jin R. and Yang L. (2006) Semi-supervised Multi-label Learning by Constrained Non-negative Matrix Factorization. In: *Proceedings of the 21st national conference on Artificial Intelligence (AAAI)*, Vol. 1, 421-426.

[58] Makadia A., Pavlovic V. and Kumar S. (2010). Baselines for Image Annotation. In: *International Journal of Computer Vision*, Vol. 90, Issue 1, 88-105.

[59] Manning C. and Schütze H. (1999) Foundations of Statistical Natural Language Processing. *MIT Press. Cambridge*.

[60] Marin-Castro H., Sucar L.E. and Morales E.F. (2007) Automatic image annotation using a semi-supervised ensemble of classifiers. In: *Progress in Pattern Recognition, Image Analysis and Applications*, LNCS Vol. 4756, 487-495.

[61] Milne D. and Witten I. H. (2008) An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia links. In: *Proceedings of the first AAAI Workshop on Wikipedia and Artificial Intelligence (WIKIAI)*, 25-30.

[62] Min J. and Jones G.J.F. (2011) External Query Reformulation for Text-based Image Retrieval. In: *Proceedings of the 18th international conference on String processing and information retrieval (SPIRE)*, LNCS Vol. 7024, 249-260.

[63] Mitra M. and Chaudhuri B.B. (2000) Information Retrieval from Documents: A Survey. In: *Information Retrieval*, Vol. 2, Issue 2-3, 141-163.

[64] Monay F. and Gatica-Perez D. (2003) On Image Auto-Annotation with Latent Space Models. In: *Proceedings of the eleventh ACM international conference on Multimedia (MM) 2003*, 275-278.

[65] Murphy K.P. (2012) Machine Learning: A Probabilistic Perspective. *The MIT Press*, 9-9.

[66] Nanni L. and Lumini A. (2013) Heterogeneous bag-of-features for object/scene recognition. In: *Applied Soft Computing*, Vol. 13, Issue 4, 2171-2178.

[67] Navarrete D.J., Morales E.F. and Sucar L.E. (2012) Unsupervised Learning of Visual Object Recognition Models. In: *Ibero-American Conference on Artificial Intelligence (IBERAMIA) 2012*, LNCS 7637, 511-520.

[68] Ngiam J., Khosla A., Kim M., Nam J., Lee H., Ng A. Y. (2011). Multimodal Deep Learning. *Proceedings of the 28th International Conference on Machine Learning (ICML)*, 1-8.

[69] Patterson G. and Hays J. (2012) SUN Attribute Database: Discovering, Annotating, and Recognizing Scene Attributes. In: *Proceeding of the 25th Conference on Computer Vision and Pattern Recognition (CVPR)*, 2751-2758.

[70] Plate T. (1991) Holographic Reduced Representations: Convolution Algebra for Compositional Distributed Representations. In: *Proceedings of the 12th International Joint Conference on Artificial Intelligence (IJCAI)*, 30-35.

[71] Putthividhya D., Attias H., Nagarajan S. S. (2010). Topic Regression Multi-Modal Latent Dirichlet Allocation for Image Annotation. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*, 3408-3415.

[72] Rapp R. (2002) The Computation of Word Associations: Comparing Syntagmatic and Paradigmatic Approaches. In: *Proceedings of the 19th international conference on Computational Linguistics (COLING)*, Vol. 1, 1-7.

[73] Rasiwasia N., Costa Pereira J., Coviello E., Doyle G., Lanckriet G. R. G., Levy R., Vasconcelos N. (2010). A New Approach to Cross-Modal Multimedia Retrieval. In: *Proceedings of the International Conference on Multimedia (MM)*, 251-260.

[74] Reshma I.A., Ullah M.Z. and Aono M. (2013) KDEVIR at ImageCLEF 2013 Image Annotation Subtask. In: *CLEF 2013 Evaluation Labs and Workshop*, Online Workings Notes. Valencia, Spain.

[75] Reza Zare M. (2013) Intelligent Methods for Automatic Classification of Medical Images. *PhD Thesis*, Faculty of Computer Science and Information Technology, University of Malaya.

[76] Sahbi H. (2013) CNRS - TELECOM ParisTech at ImageCLEF 2013 Scalable Concept Image Annotation Task: Winning Annotations with Context Dependent SVMs. In: *CLEF 2013 Evaluation Labs and Workshop*, Online Workings Notes. Valencia, Spain.

[77] Salakhutdinov R. and Hinton G. (2009). Deep Boltzmann Machines. In: *Proceedings of the international conference on artificial intelligence and statistics*. Vol. 5. No. 2, 448-455.

[78] Sánchez-Oro J., Montalvo S., Montemayor A., Pantrigo J., Duarte A., Fresno V. and Martínez R. (2013). URJC&UNED at ImageCLEF 2013 Photo Annotation Task. In: *CLEF 2013 Evaluation Labs and Workshop*, Online Workings Notes. Valencia, Spain.

[79] Serdyukov P., Murdock V. and van Zwol R. (2009) Placing Flickr Photos on a Map. In: *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 484-491.

[80] Shao Y., Zhou Y., He X., Cai D. and Bao H. (2009) Semi-Supervised Topic Modeling for Image Annotation. *Proceedings of the 17th ACM international conference on Multimedia (MM)*, 521-524.

[81] Simpson M.S., Fushman D.D., Antani S.K. and Thoma G.R. (2014) Multimodal Biomedical Image Indexing and Retrieval using Descriptive Text and Global Feature Mapping. In: *Information Retrieval*, Vol. 17, Issue 3, 229-264.

[82] Smeulders A. W.M., Worring M., Santini S., Gupta A. and Jain R. (2000) Content-Based Image Retrieval at the End of the Early Years. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, Issue 12, 1349-1380.

[83] Sorokin A. and Forsyth D. (2008) Utility data annotation with Amazon Mechanical Turk. In: *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1-8.

[84] Srivastava N. and Salakhutdinov R. (2012). Multimodal Learning with Deep Boltzmann Machines. In: *Neural Information Processing Systems (NIPS)*, 2231-2239.

[85] Sun A. and Bhowmick S.S. (2009) Image Tag Clarity: In Search of Visual-Representative Tags for Social Images. In: *Proceedings of the first SIGMM Workshop on Social Media (WSM)*, 19-26.

[86] Sun A. and Bhowmick S.S. (2010) Quantifying Tag Representativeness of Visual Content of Social Images. In: *Proceedings of the International Conference on Multimedia (MM)*, 471-480.

[87] Tang W., Cai R., Li Z., Zhang L. (2011) Contextual Synonym Dictionary for Visual Object Retrieval. In: *Proceedings of the 19th ACM international conference on Multimedia (MM)*, 503-512.

[88] Tian Dong ping (2014) A Survey of Refining Image Annotation Techniques. In: *International Journal of Multimedia and Ubiquitous Engineering(IJMUE)*, Vol. 9, Issue 3, 117-128.

[89] Tuytelaars T. and Mikolajczyk K. (2007) Local invariant feature detectors: A survey. In: *Foundations and Trends in Computer Graphics and Vision*, Vol. 3, Issue 3, 177-280.

[90] Uricchio T., Bertini M., Ballan L. and Del Bimbo A. (2013) MICC-UNIFI at ImageCLEF 2013 Scalable Concept Image Annotation. In: *CLEF 2013 Evaluation Labs and Workshop*, Online Workings Notes. Valencia, Spain.

[91] Vanegas J. A. and González F. A. (2013). Large Scale Image Indexing Using Online Non-negative Semantic Embedding. In: *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, LNCS Vol. 8258, 367-374.

[92] Verma Y. and Jawahar C.V. (2012) Image annotation using metric learning in semantic neighbourhoods. In: *Proceedings of the 12th European conference on Computer Vision*, Vol. Part III, 836-849.

[93] Villegas M., Paredes R. and Thomee B. (2013) Overview of the ImageCLEF 2013 Scalable Concept Image Annotation Subtask. In: *CLEF 2013 Evaluation Labs and Workshop*, Online Workings Notes. Valencia, Spain.

[94] Vogel J. and Schiele B. (2007) Semantic Modeling of Natural Scenes for Content-Based Image Retrieval. In: *International Journal of Computer Vision (IJCV)*, Vol. 72, Issue 2, 133-157.

[95] Xing E., Yan R. and Hauptmann (2005) Mining Associated Text and Images with Dual-Wing Harmoniums. *Proceedings of the Twenty-First Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 633-641.

[96] Yavlinsky, A., Schofield E., Rüger S. (2005) Automated Image Annotation Using Global Features and Robust Nonparametric Density Estimation. In: *Image and Video Retrieval*, LNCS Vol. 3568, 507-517.

[97] Ye M., Yin P., Lee W.C. and Lee D.L. (2011) Exploiting Geographical Influence for Collaborative Point-of-Interest Recommendation. In: *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, 325-334.

[98] Zha Z.J., Yang L., Mei T., Wang M. and Wang Z. (2009) Visual Query Suggestion. In: *Proceedings of the 17th ACM international conference on Multimedia (MM)*, 15-24.

[99] Znaidia A., Le Borgne H. and Hudelot C. (2013) Tag completion based on belief theory and neighbor voting. In: *Proceedings of the 3rd ACM International conference on multimedia retrieval*, 49-56.