Categorización de imágenes mediante técnicas de minería de texto

Adrián Pastor López Monroy, Manuel Montes y Gómez, Hugo Jair Escalante y Fabio A. González

Reporte Técnico No. CCC-16-003

Febrero 2016

Coordinación de Ciencias Computacionales

INAOE

Luis Enrique Erro No. 1,

Sta. Ma. Tonantzintla,

72840, Puebla, México.

Índice

Resumen

Hoy en día existe una cantidad inmensa de imágenes capturadas a través de distintos dispositivos para dominios específicos, por ejemplo: bases de datos médicas, registros de seguridad, redes sociales, etc. En este escenario, surge la necesidad de contar con mecanismos automáticos para facilitar el análisis de dicha información. La clasificación (e.g., ordenar las imágenes de acuerdo a categorías predefinidas) y la recuperación de imágenes (e.g., a partir de una imágen consulta, recuperar las imágenes más relevantes de una colección) son problemas de interés para distintas áreas de investigación. Actualmente, diversos enfoques reportados en la literatura subyacen sobre la base del concepto de *palabra visual*, ésta es, un elemento visual representativo de un grupo de regiones visualmente similares entre si. La representación de bolsa de palabras visuales (BoVW, *Bag of Visual Words*), es uno de los enfoques más utilizados en diversas tareas de visión computacional de alto nivel (HLCV, *High Level Computer Vision*). La BoVW es un histograma de la ocurrencia de las palabras visuales en cada imagen, en cierta forma análoga a la representación bolsa de palabras (BoW, *Bag-of-Words*) utilizada en minería de texto. Al igual que la BoW, la representación BoVW cuenta con algunas limitaciones, por ejemplo: no integra información espacial entre las palabras visuales (e.g. información contextual de otras palabras), no involucra información más allá de la ocurrencia de las palabras visuales (e.g., información semántica), produce alta dimensionalidad y una alta dispersión de la información en la representación. En esta propuesta de investigación se pretende tomar ventaja de la información contextual (e.g., espacial, secuencial) y de alto nivel (e.g., semántica) existente entre las palabras visuales, utilizando enfoques inspirados en el procesamiento del lenguaje natural (NLP, *Natural Language Processing*). El uso de métodos provenientes del NLP para la explotación de este tipo de información hace posible capturar la información contextual de los elementos visuales a través de distintos enfoques, por ejemplo, mediante representaciones basadas en n-gramas de palabras visuales, análisis de la distribución de palabras visuales, y esquemas de pesado de términos que podrían ayudar a enfatizar el contexto en el que ocurren las palabras visuales (o un conjunto de éstas), resaltando aquellos elementos que sean más descriptivos o discriminativos. Asimismo, la información de alto nivel podría ser extraída por medio de la adecuada adaptación de métodos de análisis semántico de NLP y la definición de estructuras jerárquicas con

relaciones del tipo *is-a*. El uso de enfoques de NLP puede ayudar a modelar adecuadamente todo este tipo de información contextual y de alto nivel, pero plantea diversas dificultades no triviales de resolver, por ejemplo: i) la definición de atributos visuales que puedan ser análogos a los atributos textuales y por lo tanto adecuados para el uso de enfoques de NLP, ii) la definición de estrategias adecuadas para interpretar las imágenes; en minería de textos, los documentos se pueden interpretar en una sola dirección espacial, mientras que las imágenes recaen en un plano 2D en el que no existe una forma específica de interpretarlas, iii) la forma de extraer información de alto nivel (semántica) a partir de la posición espacial de las palabras visuales, iv) la manipulación conjunta de la información contextual y de alto nivel para mejorar la clasificación y recuperación de imágenes. Bajo este escenario, es necesario definir la manera adecuada para aprovechar la información contextual y de alto nivel mediante los enfoques de NLP. En este documento también se presenta, a través de los resultados preliminares, evidencia de que la investigación propuesta es factible de realizar. Para esto, se expone el uso de lo que sería la extensión natural para aprovechar parte de la información contextual en la BoVW; la utilización de n-gramas visuales como atributos para un clasificador. Las representaciones basadas en n-gramas son muy populares en el campo de NLP, en particular dentro del área de minería de texto y recuperación de información. Para mostrar el beneficio de los n-gramas como atributos, el método es evaluado en la tarea de clasificación, y como una aplicación inicial, se utiliza una colección compuesta por imágenes de histopatología.

Palabras clave: recuperación de imágenes, clasificación de imágenes, análisis de imágenes médicas, palabras visuales, relaciones espaciales

1. Introducción

Hoy en día existe una enorme cantidad de imágenes disponibles a través de distintos medios. Probablemente debido al mayor acceso y disponibilidad de dispositivos que facilitan la captura de imágenes. Sin embargo, en muchas situaciones toda esta información es poco útil si no se cuenta con las herramientas apropiadas para su análisis. En este sentido, la clasificación y recuperación de imágenes son dos de las tareas más importantes para la organización y la explotación de la información visual en distintas áreas. El enfoque general consiste en representar las imágenes a través de vectores de características visuales y utilizar métodos estándar de aprendizaje supervisado para construir modelos de clasificación y recuperación.

La representación de imágenes es uno de los procedimientos clave para una exitosa clasificación y recuperación. Actualmente uno de los enfoques más ampliamente utilizados en tareas de visión computacional de alto nivel (HLCV, *High Level Computer Vision*) es la bolsa de palabras visuales (BoVW, *Bag of Visual Words*). La BoVW podría ser vista como una analogía de la representación de bolsa de palabras (BoW, *Bag of Words*) (McCallum y Nigam, 1998; Joachims, 1998) comúnmente utilizada en tareas de clasificación de texto y recuperación de información (ver por ejemplo en, (Turney y P., 2010)). Bajo la BoW se construyen vectores que representan a los documentos, el tamaño de los vectores es igual al número de palabras en el vocabulario de la colección de documentos; cada elemento del vector indica la presencia o ausencia de cada término en el documento. De manera similar, en tareas de HLCV que usan el enfoque de BoVW, un vocabulario de palabras visuales es generado (normalmente agrupando vectores de características visuales qué representen regiones de las imágenes y tomando el centroide de cada grupo como una palabra visual) para después representar las imágenes por histogramas de la ocurrencia de las palabras visuales en las imágenes. El enfoque de BoVW, ha sido exitosamente empleado en distintas tareas de HLCV. Por ejemplo, para categorización de imágenes (Cruz-Roa, Caicedo, y González, 2011; Cruz-Roa, Díaz, y cols., 2011; Díaz y Romero, 2012), clasificación de texturas y objetos (Zhang y cols., 2007), recuperación de vídeo (Sivic y Zisserman., 2003), recuperación de imágenes (Tirilly y cols., 2009a), reconocimiento de la actividad humana (H. Wang y cols., 2009), etc.

A pesar del hecho de que la BoVW es ampliamente utilizada, ésta tiene algunas limitaciones, una de las más importantes de esta representación es que la BoVW ignora las relaciones espaciales

entre las palabras visuales (de alguna forma heredada de la representación BoW tradicional). El contexto (tomar ventaja de la información espacial de los elementos) ha probado ser un elemento útil para aumentar el rendimiento de distintas tareas en HLCV (ver e.g., (Galleguillos y Belongie, 2010; Lazebnik y cols., 2006)). En este sentido, la información contextual entre las palabras visuales podría ser capturada a través de la adaptación de distintos enfoques de NLP, por ejemplo; los n-gramas, esquemas de pesado (funciones de ponderación para los elementos), análisis de la distribución de elementos, etc. Además de la información contextual, existe información de alto nivel que comúnmente es considerada en minería de texto. De manera general, ésta consiste en ir más allá de la ocurrencia de los términos, por ejemplo, para desarrollar análisis que construyan atributos conceptuales (e.g., atributos que representen relaciones entre términos y categorías) o expandir la información a través del uso de redes semánticas (e.g., relaciones jerárquicas entre los términos). Dado este escenario, resulta atractiva la idea de extender la representación BoVW para aprovechar el uso de la información contextual y de alto nivel. En específico, se tiene la hipótesis de que algunos métodos de procesamiento del lenguaje natural (NLP, *Natural Language Processing*) que hacen uso de la información contextual y de alto nivel, podrían dar pie a nuevos métodos y representaciones en HLCV, los cuales mejoren el rendimiento de los métodos basados en palabras visuales. De esta forma, el interés de esta investigación recae en una temática relativamente joven; la intersección de las áreas de NLP y HLCV, donde la principal contribución consistirá en el desarrollo de métodos novedosos y competitivos para clasificación y recuperación de imágenes que subyacen en los conceptos de las representaciones basadas en palabras visuales, pero que tomen ventaja de la información contextual y de alto nivel a través de métodos inspirados en el NLP.

Para desarrollar un enfoque exitoso se pondrá especial atención en la adecuada adaptación de algunas de las técnicas que han probado ser de alta utilidad en NLP. Sin embargo, para tener una idea de sí algunos de los mejores enfoques del área de NLP (e.g., representaciones textuales, pesado de términos, distintos tipos de atributos) tienen oportunidad de extender y encontrar analogías con los enfoques basados en palabras visuales, es natural comenzar por explorar y traer algunas de las ideas más básicas e intuitivas del NLP. Un ejemplo de ello son los n-gramas, secuencias de n palabras, los cuales han probado ser de gran utilidad en tareas de categorización de texto (Tan y cols., 2002). Como avances en la investigación propuesta, mostramos la mejora de la BoVW tradicional al considerar n-gramas de palabras visuales.

El resto de este documento se encuentra organizado de la siguiente forma. En la Sección 2 se revisa el trabajo relacionado con esta investigación. Posteriormente en la Sección 3 se describe la problemática. La Sección 4 presenta las preguntas de investigación, los objetivos y las principales contribuciones de esta propuesta. La Sección 5 muestra la metodología a seguir en esta investigación. Luego, en la Sección 6 mostramos de forma general el plan de trabajo para los siguientes dos años y medio. Por último, en la Sección 7 delineamos el trabajo realizado y los resultados alcanzados hasta el momento.

2. Antecedentes

En está sección se realiza una revisión del trabajo relevante para esta investigación. Se comienza con la discusión de enfoques que recaen en el empleo de palabras visuales en la literatura de HLCV; posteriormente se analizan los trabajos que toman ventaja de la información espacial y de alto nivel bajo la formulación de palabras visuales.

2.1. Palabras visuales

Las ideas claves del concepto de palabras visuales fueron introducidas por Sivic y Zisserman (2003) para hacer frente al problema de recuperación de video (Sivic y Zisserman., 2003). Gracias al buen rendimiento en esta tarea, el enfoque de palabras visuales rápidamente ganó popularidad y su utilización comenzó a propagarse hacia otros campos de HLCV tales como: recuperación de imágenes (Csurka y cols., 2004; Tirilly y cols., 2009a), clasificación de texturas y objetos (Zhang y cols., 2007), reconocimiento de actividades humanas (H. Wang y cols., 2009), filtrado de imágenes de adultos (Deselaers y cols., 2008), reconocimiento de categorías de escenas naturales (Lazebnik y cols., 2006; Boiman y cols., 2008), y otros trabajos más recientes en clasificación de imágenes médicas (J. Wang y cols., 2011; Avni y cols., 2011; Cruz-Roa, Caicedo, y González, 2011; Cruz-Roa, Díaz, y cols., 2011; Díaz y Romero, 2012).

Parte del éxito de este enfoque puede ser fácilmente explicado a través de una analogía con la BoW utilizada en tareas de clasificación de texto (McCallum y Nigam, 1998; Joachims, 1998). En este contexto, regiones de la imagen representadas por descriptores visuales, juegan el rol de las palabras que pueden ser altamente discriminativas para la identificación de una clase/tema particular (Zhang y cols., 2007). Por ejemplo, en reconocimiento de objetos, una llanta (o parte de ésta)

podría ser altamente predictiva para la categoría carro.

En el estado del arte existen diversas formas de obtener palabras visuales. Sin embargo, la mayoría de los enfoques siguen una estrategia similar a la mostrada en la Figura 1. De manera general, en la Figura 1 se observan tres etapas:

1. Un conjunto de regiones/partes/puntos de interés son extraídos a partir de las imágenes; estas regiones son representadas por descriptores visuales (vectores de características).

2. Los vectores de características son agrupados, y los centroides del proceso de agrupación son considerados como las palabras visuales, las cuales son posteriormente utilizadas para la construcción de un vocabulario visual (también llamado diccionario visual o *codebook*).

3. Las palabras visuales son utilizadas como atributos para representar a las imágenes (e.g., a través de histogramas de las palabras visuales que cada imagen contiene).

Para conocer si una imagen contiene o no una palabra visual, normalmente se mide la distancia euclidiana de cada característica visual de la imagen (e.g., regiones de la imagen representadas por algún descriptor) contra cada elemento (el centroide de cada grupo) del vocabulario visual, y se asigna la palabra del centroide más cercano. Por ejemplo, si la característica visual v_1 de la imagen está más cercana al centroide del grupo $word_1$ del vocabulario visual, entonces se dice que hemos encontrado una ocurrencia de la palabra visual $word_1$.

2.2. Enfoques de NLP aplicados a palabras visuales

Tal como ya se ha mencionado, una limitación del enfoque tradicional de BoVW es que pasa por alto la información contextual que puede existir entre las palabras visuales. Este problema ha sido una de las principales motivaciones para algunos trabajos en el estado del arte. Por ejemplo, Jamieson et al. (2007) representó algunas relaciones espaciales mediante grafos de palabras visuales con el objetivo de describir logotipos en fotografías relacionadas con deportes. Algunos otros trabajos han traído ideas más relacionadas con el NLP. Por ejemplo, en recuperación de imágenes, Zheng et al. (2006) propuso la idea de construir *frases visuales* construyendo parejas de regiones claves (también llamadas *keypoints*) que estuvieran cercanas o traslapadas con otras (de acuerdo a un umbral de distancia). Dado que lo último implica probar todos los *keypoints* de forma uno-contra-todos (para seleccionar las más cercanas de acuerdo al umbral de frecuencia), lo que hacen

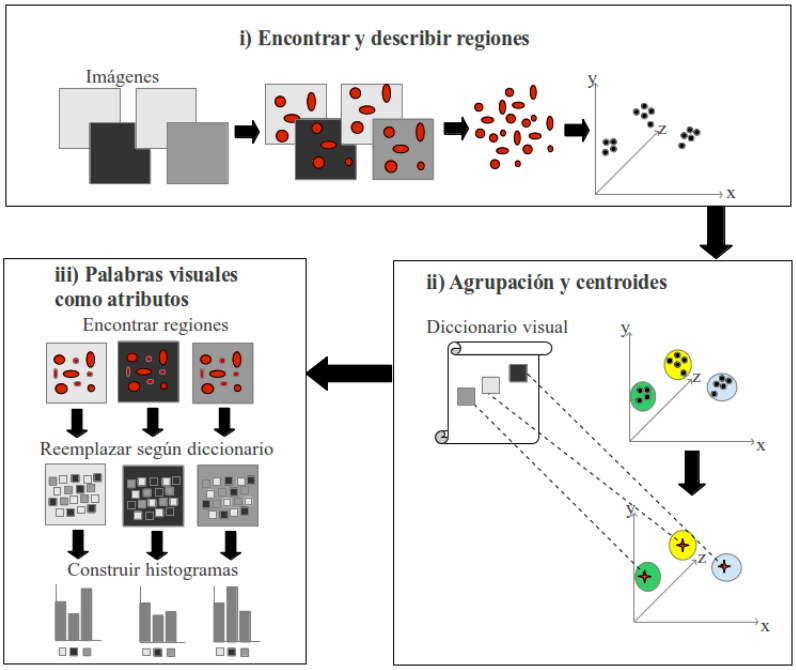

Figura 1. Enfoque general para extraer palabras visuales.

es probar esto solamente para los *keypoints* más frecuentes del conjunto de imágenes. En otros trabajos, Yuan et al. (2007, 2011) tomaron ventaja del uso del algoritmo de k vecinos más cercanos para agrupar palabras visuales y construir *frases visuales* de distintas longitudes. En minería de datos para vídeo, las frases visuales también han sido utilizadas para identificar a personajes y objetos principales en un vídeo por medio de la agrupación de escenas de vídeo (Sivic y Zisserman, 2004). En otro trabajo, Quack et al. (2007) ha explorado conjuntos locales (en un vecindario) de palabras visuales para detectar características de clases de objetos distintivas y frecuentes, esto incluso ofrece la opción de utilizar el método para reconocimiento de objetos o como selector de características. Otros enfoques han utilizado Modelos de Lenguaje (LMs, *Language Models*) con el objetivo de capturar información contextual. Los LMs son un enfoque popular utilizado en NLP para modelar secuencias de palabras. Los trabajos reportados que utilizan LMs para tareas de HLCV llevan a cabo diversos pasos antes del entrenamiento de los modelos (Wu y cols., 2007), por ejemplo: consideran el uso de co-ocurrencias e información de proximidad de palabras visuales en un vecindario. Lo último es debido a que los LMs necesitan "leer" las palabras visuales en alguna dirección. Por ejemplo, Tirilly et al. (2008), utilizó PCA (por, *Principal Component Analysis*) para proyectar descriptores visuales en ejes de dirección particulares, para después inducir la dirección

en que una secuencia de palabras visuales debe ser leída. Las secuencias de palabras visuales en las imágenes son categorizadas por medio de un clasificador basado en LMs. Este clasificador construye un LM para cada clase utilizado los documentos de entrenamiento. Para la evaluación, en cada imagen se mide la probabilidad de pertenencia a los LMs, y se predice la clase más probable.

La información contenida a través de las palabras visuales no solo se limita a la información contextual. En este sentido, diversos trabajos han explorado ideas inspiradas en enfoques de minería de textos que van más allá de contabilizar las palabras visuales y extraer su información espacial. Por ejemplo, Tirilly y cols. (2009) , utiliza n vectores de palabras visuales ponderadas para recuperación de imágenes. En estos vectores, el valor de cada elemento corresponde a la importancia de cada palabra visual en la imagen. Para lograr esto, aplica esquemas de pesado regularmente utilizados en recuperación de información (e.g., BM25, TF*IDF), y muestra que una buena elección del esquema de pesado para dominios específicos, puede mejorar el rendimiento de los enfoques hasta en un 10 %. Algunos otros trabajos han tomado ventaja de enfoques para generar atributos de alto nivel (Sivic y cols., 2005; Cao y Fei-Fei, 2007; X. Wang y Grimson, 2007). Por ejemplo, Sivic y cols. (2005) para la discriminación de objetos en imágenes, determinan conceptos semánticos de palabras visuales y su localización en las imágenes a través de modelos estadísticos de texto; análisis semántico latente probabilista (pLSA, *probabilistic Latent Semantic Analysis*) y asignación Dirichlet latente (LDA). En análisis de texto, el pLSA y LDA son utilizados para descubrir tópicos en una colección de documentos a través de la representación de BoW. La utilización de enfoques que descubren conceptos de alto nivel, hace posible descubrir categorías de objetos como si fueran tópicos, en donde imágenes con alta variabilidad visual (conteniendo distintos tipos de objetos e.g., carros, motos, etc.) pueden ser representadas como una combinación de distintos tópicos.

3. Problemática

En los trabajos anteriores los autores han propuesto interesantes extensiones a las representaciones basadas en palabras visuales, logrando reportar mejoras en el rendimiento de la clasificación y recuperación sobre la representación estándar de BoVW. Sin embargo, algunas de estas propuestas no necesariamente corresponden a la forma en la que las secuencias de palabras son procesadas en NLP para el incremento del rendimiento. Por ejemplo, los clasificadores basados en LMs y medidas de pertenencia (e.g., perplejidad, probabilidad) son en la actualidad raramente utilizados para la

categorización de texto. En vez de ello, el estado del arte ha mostrado que en muchas tareas (e.g., la clasificación temática, recuperación de información) utilizar como atributos las secuencias de palabras con enfoques de aprendizaje automático resultan tener mejor rendimiento (e.g., *Support Vector Machine, Random Forest, etc.*). Por otro lado, la obtención de otro tipo de información diferente a la espacial (e.g., esquemas de pesado, conceptos semánticos) no ha sido explorada ampliamente. Por ejemplo, en los esquemas de pesado para palabras visuales, aunque resulta interesante la mejora obtenida en las colecciones, ésta decae considerablemente en conjuntos de datos que presenten alta variabilidad visual (Tirilly y cols., 2009b). Por otra parte, el problema de generar automáticamente atributos de alto nivel (e.g., conceptos semánticos) ha sido ampliamente estudiado para el caso de clasificación y recuperación de documentos, pero existe poca literatura de HLCV que explote otras representaciones semánticas (diferentes a pLSA y LDA) inspiradas en NLP. En este contexto, resulta atractiva la idea de generar nuevos métodos que mejoren la BoVW, a través de la experiencia mostrada en el área de minería de textos y enfoques basados en BoW.

En esta propuesta de investigación, se tiene la hipótesis de que la información contextual y de alto nivel entre las palabras visuales, si es explotada adecuadamente, podría ser de gran utilidad para capturar patrones visuales en algunas tareas relacionadas con visión computacional (e.g., clasificación, recuperación de imágenes). Aprovechar dicha información por medio de la utilización de técnicas inspiradas en el NLP es algo que ha llamado la atención de la comunidad científica. Es por ello que, en esta propuesta de investigación se tiene interés en los enfoques del NLP que pudieran ser combinados con enfoques de HLCV para evaluar, encontrar y producir automáticamente características visuales más informativas.

El uso de enfoques de NLP podría hacer posible explotar otro tipo de información en las palabras visuales. Por ejemplo, la información contextual local o global, que podría ser capturada a través de la adaptación de enfoques de minería de texto tales como: n-gramas (Bekkerman y Allan, 2004; Tan y cols., 2002; S. Wang y Manning, 2009), grafos de n-gramas (Giannakopoulos y cols., 2012, 2008), histogramas pesados de n-gramas locales (Lebanon y cols., 2007; Escalante y cols., 2011), colocaciones de palabras (Scott, 2001), etc. Además de información contextual, dependiendo del éxito de las características anteriores, podrían construirse otro tipo de atributos visuales de más alto nivel, tales como los conceptos visuales, los cuales podrían ser inspirados por representaciones de análisis semánticos (Deerwester y cols., 1990; Blei y cols., 2003; Lavelli y cols., 2004; Li y cols.,

2011).

4. Preguntas, objetivos y contribuciones

En la presente propuesta se pretende contestar las siguientes preguntas de investigación:

- ¿Qué enfoques de procesamiento del lenguaje natural pudieran ser aplicados utilizando la analogía palabras-textuales/palabras-visuales para obtener métodos competitivos para la clasificación y recuperación de imágenes?
- ¿De qué manera se pueden aplicar enfoques del procesamiento del lenguaje natural para tomar ventaja de la información contextual (espacial) existente entre las palabras visuales?
- A través de enfoques motivados por el procesamiento del lenguaje natural, ¿De qué manera la información de alto nivel (e.g., semántica) podría ser extraída de las palabras visuales?

4.1. Objetivo general

- Desarrollar métodos para la clasificación y recuperación de imágenes, que sobre la base del concepto de palabras visuales y la adaptación de enfoques inspirados en NLP, permitan construir representaciones que modelen la información contextual y de alto nivel para mejorar el enfoque de bolsa de palabras visuales.

4.2. Objetivos específicos

1. Desarrollar un método para la representación de imágenes que capture información contextual discriminativa a partir las palabras visuales (e.g., a través de la información espacial de los elementos, secuencias de palabras visuales, esquemas de pesado, etc.).
2. Desarrollar un método para la representación de imágenes que capture información de alto nivel discriminativa a partir las palabras visuales (e.g., a través de análisis semánticos, relaciones jerárquicas del tipo *is-a*, etc.).
3. Evaluar la utilidad de las distintas representaciones generadas en el contexto de aplicaciones a dominios específicos (e.g., clasificación y recuperación de imágenes médicas, de texturas, naturales, etc.).

4. Diseñar e implementar un método para clasificar y recuperar imágenes, que tome ventaja de la información de las diferentes representaciones generadas a partir de las palabras visuales.

4.3. Contribuciones esperadas

A través de esta investigación doctoral se espera obtener las siguientes contribuciones:

- Un conjunto de métodos para generar representaciones de imágenes inspiradas en NLP que sean de utilidad para capturar la información contextual y que extiendan la bolsa de palabras visuales.
- Un conjunto de métodos para generar representaciones para obtener información de alto nivel a partir de las palabras visuales, que en analogía con estrategias de NLP, logre capturar información más allá de la frecuencia y distribución de los atributos.
- Un estudio detallado de la utilidad en clasificación y recuperación de imágenes de las representaciones propuestas en diferentes dominios (e.g., imágenes médicas, de texturas, naturales, etc.).
- Un enfoque híbrido para clasificar y recuperar imágenes, que logre aprovechar la información conjunta de las diferentes representaciones visuales propuestas.

5. Metodología

En esta sección se explica en detalle la metodología propuesta para alcanzar los objetivos planteados. La metodología planteada consta de cinco pasos secuenciales, siendo los pasos 3 y 4 los que concentran las principales contribuciones de esta propuesta.

1. **Identificación y obtención de los conjuntos de datos.**

    1.1. **Identificar conjuntos de datos en distintos dominios de imágenes.** Estos pueden ser imágenes médicas, imágenes utilizadas para reconocimiento de objetos, imágenes utilizadas para reconocimiento de texturas, etc. El propósito es encontrar colecciones en dónde puedan existir patrones que sea posible caracterizarlos en conceptos de alto nivel (palabras visuales), pero con una o más peculiaridades desafiantes que las técnicas de NLP pudieran tratar. A continuación se muestran algunos criterios para guiar la selección de los conjuntos de datos, y se explica el escenario análogo en minería de textos que pensamos, podría motivar nuevas soluciones.

    - *Criterio de información estructural:* En el caso de las imágenes, se presenta en problemas en el que la información de estructura/contextual/espacial de los elementos visuales es de importancia. En minería de texto esta información es interesante en las tareas de clasificación y recuperación de documentos (e.g., capturar el contexto de las palabras permite tener información menos ambigua) (Tan y cols., 2002), este tipo de información es frecuentemente capturada a través de: modelos del lenguaje, n-gramas a nivel de palabras y caracteres, y esquemas de pesado de términos a través del documento.
    - *Criterio de información traslapada:* En el caso de las imágenes, se presenta en colecciones que contienen un alto traslape en el contenido visual para elementos que son de distintas categorías. En minería de texto esto es un problema común en atribución de autoría (e.g., cuando se trata de discriminar los documentos de distintos autores escribiendo sobre la misma temática) (Schler y cols., 2009; Koppel y Schler, 2003), normalmente abordado mediante la consideración de la ocurrencia de las palabras vacías (*stopwords*), y el análisis de la distribución de los elementos comunes entre todos los autores (Stamatatos, 2009).
    - *Criterio de información heterogénea:* En el caso de las imágenes, se presenta en conjuntos de datos que contienen una alta variabilidad de la apariencia visual para elementos que pertenecen a la misma clase (lo cual dificulta la tarea de extraer patrones relevantes). En minería de texto esta es una situación presente en dónde

es necesario modelar observaciones más generales de grupos de individuos escribiendo (Schler y cols., 2006). Por ejemplo, el perfilado de autores, en dónde el interés es discriminar entre el género, la edad, lenguaje nativo, grupos sociales, etc. normalmente este problema es abordado a través de técnicas similares a las de atribución de autoría, y representaciones de alto nivel basados en frecuencia de términos, co-ocurrencias de términos, y conceptos semánticos (Argamon y cols., 2009; Koppel y cols., 2002).

1.2. **Adquirir y construir los conjunto de datos adecuados para la evaluación de los métodos.** Este paso implica hacer el tratamiento y procesamiento necesario a las imágenes para una adecuada evaluación de los métodos.

2. **Analizar y desarrollar métodos para la extracción de palabras visuales.** Debido a que existe una gran variedad de métodos para llevar a cabo la extracción de las palabras visuales, es importante encontrar y adaptar un método adecuado para obtener una mejor analogía entre palabras textuales y palabras visuales. En este sentido, la identificación de las regiones de la imagen juegan un rol importante en la construcción de las palabras visuales.

   2.1. **Extracción de regiones a través de mallas:** En este paso se considera el desarrollo de un enfoque de extracción de regiones de la imagen a través de una malla con tamaño de parches uniformes (Winn y cols., 2005; Nowak y cols., 2006). Se tiene la hipótesis de que este enfoque es de utilidad dado que la analogía entre texto e imagen cuadriculada podría ser más natural.

   2.2. **Extracción de regiones a través de puntos clave:** Consiste en solamente extraer regiones relevantes en la imagen (e.g., esquinas, texturas particulares) a través de un algoritmo de visión computacional (Lowe, 2004).

3. **Proponer un conjunto de nuevas representaciones basadas en palabras visuales y métodos inspirados en NLP, que puedan ayudar a capturar información contextual.** Para esto es necesario determinar qué métodos de NLP podrían ser adaptados para retener información contextual visual relevante. En este escenario, consideramos a los siguientes enfoques como los mejores candidatos para la extracción de información contextual:

3.1. **Secuencias de palabras visuales**: La idea general consiste en extraer secuencias de palabras visuales de forma similar en como han probado ser útiles en el área de minería de texto (Tan y cols., 2002), pero tomando en cuenta las particularidades del dominio de las imágenes. Actualmente los n-gramas de palabras (secuencias consecutivas de n palabras) son uno de los enfoques más utilizados. Entre las consideraciones a tomar en cuenta para extraer este atributo se encuentran:

- La forma adecuada de extraer estas secuencias de palabras visuales (dado que las imágenes recaen en un plano 2D).
- La forma de interpretarlos (en texto es sólo de izquierda a derecha) para obtener una correcta contabilización y hacerlos más tolerantes al problema de rotación.

Mejorar el rendimiento (en clasificación y recuperación) de la BoVW a través de un enfoque inspirado en el modelo de n-gramas, puede motivar a extraer secuencias de elementos más elaboradas. Por ejemplo, el uso de patrones de secuencias frecuentes maximales, esto es, secuencias de elementos frecuentes que no están contenidas en otras secuencias de elementos (García-Hernández y cols., 2004, 2006)).

3.2. **Histogramas de n-gramas visuales localmente pesados**: Este método ha demostrado ser valioso para diversas tareas (Lebanon y cols., 2007), siendo la atribución de autoría una de las más interesantes para minería de textos (Escalante y cols., 2011). Este enfoque consiste en extraer y construir histogramas de n-gramas de un documento, pero asignándoles un peso de acuerdo a la posición en dónde ocurren. De esta forma, se puede enfatizar mejor distintas partes interesantes de un documento según la tarea. Por ejemplo, en atribución de autoría, resulta relevante asignarles diferente importancia a los términos que cada autor prefiere según ocurran en el inicio, desarrollo o conclusión de un documento. De manera similar, en el dominio de las imágenes, se podría asignar mayor importancia a ciertas regiones de las imágenes según la tarea. Por ejemplo, en el dominio médico, en el análisis de imágenes de enfermedades neurodegenerativas, podría ser útil enfatizar de distinta forma las regiones que presenten actividad de las neuronas. Otro ejemplo es en imágenes de histopatología, cuando una enfermedad está asociada a aglomeraciones de elementos visuales específicos (células de cáncer) ,

entonces se identifican estas aglomeraciones (e.g., usando alguna medida de entropía), para después usar este enfoque para enfatizar las partes de la imagen que presenten este comportamiento. Entre las consideraciones no triviales de resolver para adaptar este enfoque, además de las anteriores heredadas por ser n-gramas, se encuentra:

- Diseñar una función de pesado para determinar la importancia de una región visual 2D.
- Desarrollar un enfoque automático para determinar las regiones de la imagen que se quieren enfatizar según la tarea.

3.3. **Grafos de n-gramas de palabras visuales:** Este tipo de representación ha sido utilizado en minería de texto para realizar resúmenes automáticos y clasificación supervisada (Giannakopoulos y cols., 2008, 2012). La idea general consiste en construir grafos de documentos en donde cada elemento textual es un nodo, y se conecta con otros nodos que representan los elementos con los que co-ocurrió, los grafos se encuentran conectados por aristas que modelan la frecuencia, co-ocurrencia y orden. Esto permite capturar el orden de la aparición de todos los elementos en el documento (la cual es ignorada por los n-gramas comunes). Un ejemplo de ello es la cadena "Yo tengo un wiki dónde escribo acerca del kiwi, amo mi wiki", en ella, las subcadenas "wiki" y "kiwi", en modelo de n-gramas de caracteres común, producen exactamente los mismos bigramas, pero en el modelo de grafos de n-gramas esto no sucede, debido a que logra modelar el documento completo y capturar estas diferencias (por que mantiene información de la co-ocurrencia de todos los n-gramas con todos) (Giannakopoulos y cols., 2008, 2012). Entre las dificultades a resolver al adaptar este enfoque se tiene:

- Realizar un análisis para determinar la forma adecuada en la que las palabras visuales deben ser interpretadas (e.g., determinar la forma en el que las palabras visuales están co-ocurriendo) para la construcción del grafo.
- Definir la noción de *stopword visual* para tomarla en cuenta o no en esta representación de grafos según la naturaleza de las imágenes. Por ejemplo, el *stopword visual* podría ser útil solo si las categorías presentan un alto traslape en el contenido visual.

El siguiente paso es complementar a los atributos que capturen la información espacial por medio de atributos de más alto nivel. Para generar estos meta-atributos se pretende generar conceptos a partir de la adaptación de enfoques de NLP (e.g., LSA). Estos atributos complementaran a los atributos calculados anteriormente, incluso usarlos como base para construir conceptos.

4. **Proponer un conjunto de nuevas representaciones basadas en palabras visuales y métodos inspirados en NLP, que puedan ayudar a capturar información de alto nivel** Para esto es necesario determinar qué métodos de NLP podrían ser adaptados para retener información de alto nivel visual relevante. En este contexto, se tienen los siguientes enfoques como posibles opciones.

   4.1. **Análisis semántico conciso (CSA, *Consice Semantic Analysis*) de palabras visuales:** A grandes rasgos el CSA en clasificación de texto consiste en tener las características textuales en vectores de términos que representen las relaciones con cada una de las categorías (la dimensionalidad está dada por el número de categorías). Usando estos vectores de término, es posible construir representaciones que indiquen las relaciones de cada documento con cada categoría (Zhixing y cols., 2010; López-Monroy y cols., 2012, 2013). De manera similar, en el dominio de las imágenes, la idea es utilizar CSA para determinar el grado de relación que cada palabra visual guarda con las categorías (algo importante desde el punto de vista de interpretabilidad de los resultados), y construir representaciones con baja dimensionalidad, densas, y más tolerantes al ruido. Entre los problemas no triviales de resolver en la adaptación de este enfoque se encuentra:

      - Determinar una función para ponderar adecuadamente la relación que una palabra visual tiene en la imagen.
      - Generar automáticamente una cantidad adecuada de atributos semánticos que permita aprovechar al máximo esta representación según la tarea.

   4.2. **Estructuras basadas en conocimiento (relaciones *is-a*):** En minería de texto existen los enfoques que usan ontologías (estructuras para la representación del conocimiento) que contienen palabras relevantes y sus relaciones dentro de un dominio. Un ejemplo de ello es *wordnet*, una base de datos léxica del inglés, que agrupa conjuntos de sinónimos

llamados *synsets*, proporcionando definiciones cortas y almacenando las relaciones semánticas entre los conjuntos de sinónimos. Algunos enfoques de NLP, a través de ciertas restricciones, han utilizado a *wordnet* como una ontología interpretando la relación entre los *synsets* como una relación de especialización entre categorías conceptuales. Tener esta jerarquía de relaciones entre palabras permite utilizarla para diferentes propósitos, tales como: desambiguación del sentido de las palabras y expansión de los términos. Esto ha mostrado ser útil en tareas como recuperación de información, clasificación automática de texto, generación de resúmenes automáticos de texto, traducción automática, entre otras. De manera análoga, se tiene la idea de que tener una estructura jerárquica de palabras visuales podría mejorar significativamente el rendimiento de las tareas de clasificación y recuperación de imágenes. Entre los problemas más desafiantes para construir este tipo de estructuras se encuentra:

- Desarrollar un método para construir una estructura jerárquica para las regiones visuales. Para lograr esto se tiene contemplado que al construir las palabras visuales, se utilice un enfoque de agrupación jerárquico *bottom-up*, también llamado agrupación aglomerativa (Murtagh, 1983). Para ello, es necesario definir una adecuada medida de distancia para dos grupos de regiones visuales. La idea es empezar con cada instancia como un grupo; luego encontrar los dos grupos más cercanos, mezclarlos, y continuar con este procedimiento hasta que sólo quede un grupo. El registro de estas combinaciones formaría una estructura de agrupación jerárquica, esto es un árbol binario, que podríamos inicialmente tomarlo como nuestra estructura de relaciones *is-a*.

5. **Desarrollar e implementar un método para llevar a cabo la clasificación y recuperación de imágenes que integre la información extraída con los enfoques de NLP (información contextual y la de alto nivel).**

   Este último paso involucra el desarrollo de un método para la clasificación y recuperación de imágenes que pueda tomar en cuenta distintos tipos de atributos para la clasificación. Por ejemplo, un algoritmo de ensamble que pueda sacar ventaja de distintos atributos como, palabras visuales, n-gramas de palabras visuales, oraciones de palabras visuales, conceptos (e.g.,

usando representaciones como CSA), y la información de la estructura jerárquica, etc. El conjunto de ensambles podrían ser combinados a través de técnicas de fusión de información como:

- *Fusión tardía*: cada conjunto de atributos, propiamente representados en un vector, es tomado para entrenar un clasificador del ensamble, o bien lanzar una consulta, ponderando y mezclando los resultados obtenidos (Kuncheva, 2005).
- *Fusión temprana*: toda la información es tomada como un solo vector para entrenar un clasificador o lanzar una consulta (Kuncheva, 2005).
- *Aprendizaje de múltiples kernels para clasificación*: cada conjunto de atributos propiamente representados en un vector, es tomado para entrenar una maquina de vectores de soporte, posteriormente los kernels son combinados en uno mismo (e.g., generalmente a través de operaciones lineales) (Gönen y Alpaydın, 2011).

6. Plan de trabajo

A continuación se presenta de forma general un plan de trabajo hasta los siguientes dos años y medio para algunas de las tareas más relevantes que se tienen planeadas hasta el momento.

Tabla 1. Plan de trabajo

Tarea	2013												2014												2015												2016				
	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5
Revisión del estado del arte																																									
Identificación y obtención del conjuntos de datos																																									
Análisis y procesamiento de los conjunto de datos																																									
Análisis y desarrollo de métodos para la extracción de palabras visuales																																									
Pruebas de n-gramas visuales sobre el conjunto de datos																																									
Análisis de resultados preliminares																																									
Elaboración de artículo																																									
Elaboración de la propuesta																																									
Defensa de la propuesta													X																												
Experimentos y análisis de métodos para extraer información contextual																																									
Escribir artículo de revista																																									
Segunda evaluación																									X																
Experimentos y análisis de métodos para extraer información de alto nivel																																									
Llevar a cabo la clasificación de imágenes a través de una fusión multimodal																																									
Escribir un segundo artículo de revista																																									
Escritura del documento de tesis																																									
Tercera evaluación																																						X			
Experimentación complementaria																																									
Entrega de la tesis																																							X		
Revisión y correción del documento de tesis																																									
Defensa de tesis																																									X

Actividad realizada.
Actividad pendiente.
X Entrega

7. Trabajo realizado y resultados preliminares

El trabajo realizado hasta hoy consiste en lo siguiente:

1. Identificación y obtención del primer conjunto de datos (**parte del primer paso de la metodología**). Para llevar a cabo este paso, se colaboró con investigadores en el área de análisis de imágenes de la Universidad Nacional de Colombia. Las imágenes de histopatología, por la razones explicadas en la Sección 7.1, fueron seleccionadas cómo un primer conjunto de datos para probar las ideas iniciales de esta propuesta.

2. Desarrollo del método para la obtención del diccionario de palabras visuales a través de mallas (**parte del segundo paso de la metodología**). Parte de las motivaciones específicas de usar un enfoque basado en una cuadricula son:

   - ⋆ Permite obtener fácilmente la información espacial entre las regiones de la cuadricula.
   - ⋆ Permite llevar a un descriptor gran parte de las regiones de una imagen. Otros esquemas de extracción de regiones visuales solo obtienen ciertos elementos con bordes o esquinas interesantes, e ignoran regiones uniformes y muy frecuentes entre todas las clases de objetos. Desde el punto de vista de minería de texto, esto puede ser de útilidad en algunas situaciones. Por ejemplo, se podría definir el concepto de *palabra vacía*, la cual en muchas tareas de clasificación de texto es filtrada por su escasa información discriminativa (e.g., en la clasificación temática), pero en muchas otras (cuando se tiene documentos de la misma temática) es clave por su información descriptiva para detectar elementos relacionados con el estilo (e.g., atribución de autoría, detección del perfil de autores) (Stamatatos, 2009; Schler y cols., 2009; Koppel y Schler, 2003). Por otro lado, en el ámbito de visión, si estamos tratando con imágenes muy parecidas (e.g., detección de tipos de hojas o plantas) o provenientes de una misma fuente (e.g., biopsias de la piel), la definición de una *palabra visual vacía* resulta muy atractiva desde el punto de vista análogo a minería de texto (cuando se trata con elementos de temáticas muy similares).

3. Desarrollo de la bolsa de n-gramas de palabras visuales como una primera aproximación para la captura de la información contextual (**parte del tercer paso de la metodología**). Parte de

las actividades realizadas para la realización de este punto son:

- ⋆ Evaluación de la utilidad de los n-gramas de palabras visuales.
- ⋆ Reportar resultados en artículo de conferencia arbitrada.

7.1. Identificación y obtención del primer conjunto de datos

La mayoría del trabajo en BoVW ha sido sobre el uso de imágenes naturales en dónde es importante la apariencia (e.g., paisajes, texturas) para identificar la clase. En este sentido, para nuestra evaluación, y como una primera aplicación para los enfoques que se utilizarán, se ha decidido abordar la clasificación automática de imágenes de histopatología. Las imágenes histopatológicas presentan algunas peculiaridades que las distinguen de análisis de las imágenes naturales, entre las que se encuentra:

- Contenido visual heterogéneo y rico.
- Alta variabilidad de la apariencia visual dentro de la misma clase.
- Mezclas complejas de patrones estructurales locales y globales.

En particular, una representación BoVW estándar asumiría que existen patrones (palabras visuales) que pueden caracterizar conceptos de alto nivel en las imágenes; lo cual no es necesariamente cierto para imágenes histopatológicas en dónde la información estructural puede ser relevante (Cruz-Roa, Díaz, y cols., 2011). Es por ello que, en esta experimentación preliminar consideramos la clasificación automática de imágenes de histopalogía de acuerdo a las estructuras de tejidos (sanos o patológicos) que pueden ser reconocidos por una inspección visual de un patólogo experto (ver Figura 2). Esas imágenes son particularmente complejas, y su clasificación está relacionada con lesiones patológicas, morfológicas y características de estructura variable, las cuales engloban una mezcla compleja de patrones visuales para decidir acerca de la presencia de la enfermedad.

Para evaluar los enfoques iniciales de esta propuesta se considera un conjunto de imágenes de histopatología, etiquetadas por un patólogo, que describen la presencia de características visuales estructurales, morfológicas y tejidos patológicos (Díaz y Romero, 2012; Cruz-Roa, Díaz, y cols., 2011). Las instancias corresponden a imágenes de color RGB con un aumento de 10X teñidas con

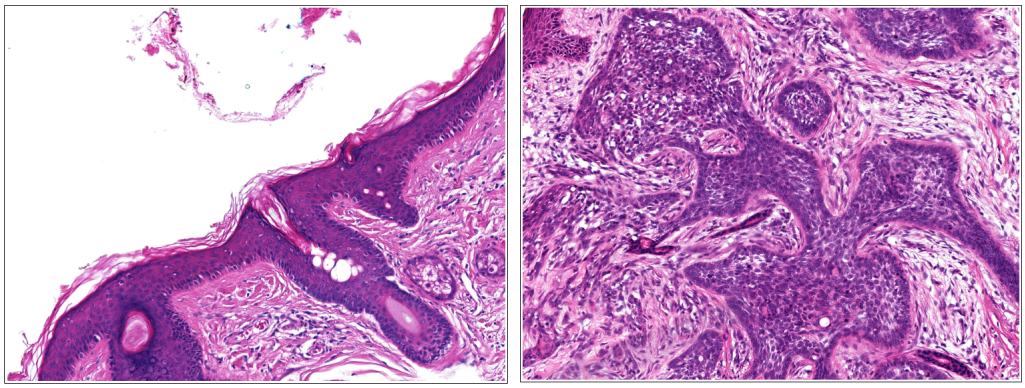

Figura 2. Ejemplo de imágenes de histopatología. Biopsias de la piel con tejido sano (*ephitelium*) y patológico (*morpheafor basal-cell carcinoma*), izquierda y derecha respectivamente, utilizados para el diagnóstico de *basal-cell carcinoma*.

hematoxilina-eosina (H&E) de tejidos de la piel diagnosticados como sanos (colágeno, la epidermis, folículos pilosos, glándulas ecrinas, glándulas sebáceas y el infiltrado inflamatorio), o por la presencia de células basales carcinoma (BCC, *Basal-Cell Carcinoma*) (que es el único relacionado con diagnóstico de cáncer).

Para este estudio se toma un subconjunto del corpus original de histopatología (Díaz y Romero, 2012), el cual está compuesto por 1,417 imágenes de histopatología de 300X300 pixeles, dónde cada una puede pertenecer a una o más de las 7 diferentes categorías relacionadas con estructuras específicas y morfológicas de tejidos sanos o patológicos para el diagnóstico de BCC. En la Tabla 2 se muestra la distribución de imágenes de histopatología para cada una de las siete categorías.

Imagen histopatológica	**positivas**	**negativas**
1. Células basales carcinoma	518	899
2. Colágeno	1238	179
3. Epidermis	147	1270
4. Folículo piloso	118	1299
5. Glándulas ecrinas	126	1291
6. Glándulas sebáceas	136	1281
7. Infiltrado inflamatorio	99	1318

Tabla 2. Distribución por categoría de las imágenes de histopatología.

7.2. Captura de la información contextual en las imágenes a través del uso de n-gramas de palabras visuales.

Como una primera aproximación a la explotación de la información contextual contenida en las imágenes, se propone dar un paso más allá de la bolsa de palabras visuales (BoVW, *Bag-of-Visual Words*) estándar: explorar la utilización de la bolsa de n-gramas de palabras visuales (BoNVW, *Bag-of-Visual Ngrams*). Los n-gramas son secuencias de n palabras, las cuales han sido ampliamente utilizados en NLP, en particular dentro del área de categorización de texto y recuperación de información (Tan y cols., 2002; Bekkerman y Allan, 2004; S. Wang y Manning, 2009). Este tipo de representaciones pueden capturar patrones de palabras compuestas, e.g., *estados-unidos*, *muy-bien*, *palabra-visual*, etc. Así pues, de manera similar se propone construir diccionarios visuales (también llamados *codebooks*) de n-gramas de palabras visuales (secuencias multi-direccionales de palabras visuales) para después representar las imágenes mediante la utilización de BoVW. La hipótesis es que esta representación puede capturar patrones espaciales informativos que podrían ayudar a mejorar el rendimiento de la clasificación en tareas de categorización de imágenes.

Las principales contribuciones de este experimento son dos. Primero, se introduce un efoque adecuado para el empleo de los n-gramas bajo el marco de trabajo de la BoVW en clasificación de imágenes; dónde los n-gramas son utilizados como atributos para un modelo de clasificación. Segundo, mostramos que la BoVN puede superar el rendimiento del enfoque BoVW en la tarea de categorización de imágenes de histopatología.

A continuación se explica en detalle el enfoque de bolsa de n-gramas visuales (BoVN, *Bag-of-Visual N-grams*). En la Figura 3 se muestra el proceso general para construir la BoVN. En el primer paso se toman todas la imágenes (de entrenamiento) y se extraen las palabras visuales a través del procedimiento delineado en la Sección 7.2.1. Para propósitos de esta investigación, dividimos las imágenes mediante una malla y extraemos características de cada parche (Cruz-Roa, Caicedo, y González, 2011). En un segundo paso, cada parche de cada imagen es reemplazado por la palabra visual más cercana generada en el paso uno (Sección 7.2.1). El tercer paso involucra la extracción de los n-gramas con el objetivo de construir nuestro diccionario de n-gramas visuales (explicado en la Sección 7.2.2). En el cuarto y último paso se hace uso del diccionario de palabras visuales más el diccionario de n-gramas de palabras visuales, con el objetivo de obtener un diccionario visual

final. Posteriormente se utiliza el diccionario visual final para construir histogramas de n-gramas de palabras visuales para cada imagen. Cada uno de estos pasos se describe a detalle en el resto de esta sección.

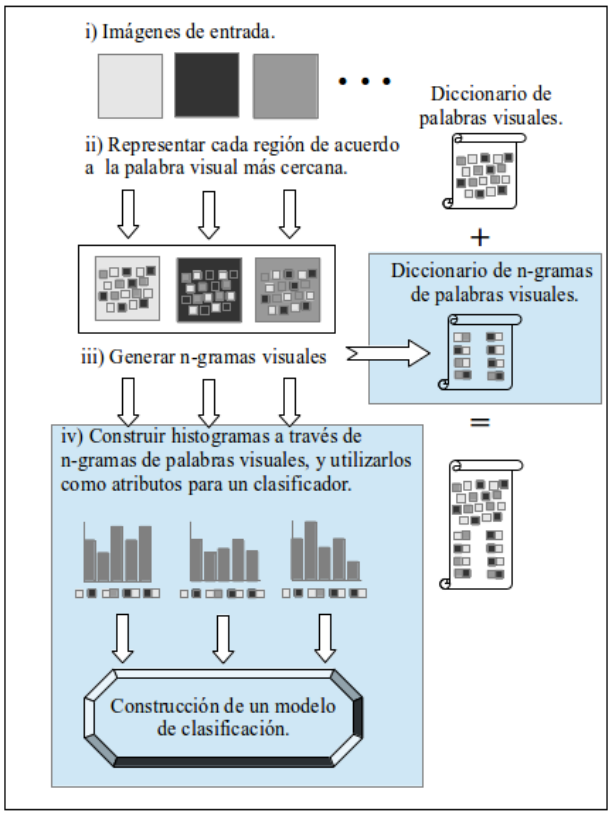

Figura 3. Representación de la imagen a través de la bolsa de n-gramas visuales.

7.2.1. Desarrollo del método para la obtención del diccionario de palabras visuales.

En la Figura 4, se muestra el procedimiento para la extracción de las palabras visuales para una colección de imágenes utilizando un enfoque de típico de BoVW. Se inicia por la obtención de pequeños parches de las imágenes. Para esto, se hace uso de un extracción basada en una malla de elementos uniformes. Esto se hace dividiendo a las imágenes a través de la malla, y tomando cada elemento de la malla como un parche de tamaño fijo, ver paso ii) en la Figura 4.

El siguiente paso consiste en representar cada parche extraído por medio de un conjunto de características. De entre la amplia variedad de descriptores de imágenes en la literatura, se utiliza la transformada discreta del coseno (DCT, *Discrete Cosine Transform*) aplicada a cada canal del espacio de colores RGB para cada parche. El descriptor es construido mezclando los 64 coeficientes de

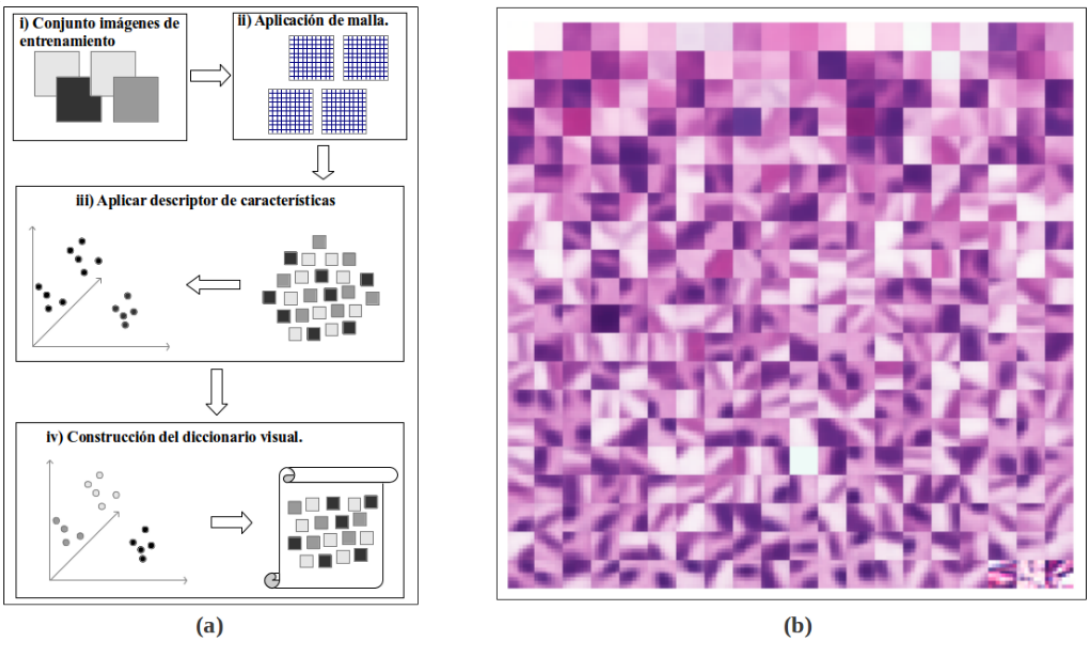

Figura 4. (a) El proceso de construcción de un diccionario de palabras visuales. (b) Ejemplo de un diccionario de palabras visuales generado.

cada uno de los tres canales. Se consideraron estas características debido a que en estudios previos éstas han superado a representaciones alternativas (incluyendo características SIFT y parches en su forma original) (Cruz-Roa, Caicedo, y González, 2011; Cruz-Roa, Díaz, y cols., 2011; Díaz y Romero, 2012). Sin embargo, otro tipo de descriptores podrían ser considerados en el futuro. Este proceso es el tercer paso en la Figura 4.

El cuarto y último paso es el proceso de construcción del diccionario visual. El diccionario es construido por medio de la agrupación de los descriptores de parches extraídos de la colección de imágenes. Esto se hace utilizando un simple algoritmo de K-*Means* con k=400, la elección de la k está motivada por un estudio previo sobre este mismo tipo de imágenes (Cruz-Roa, Díaz, y cols., 2011). En este proceso, todos los descriptores de parches similares en el conjunto de entrenamiento son agrupados. De esta forma, el algoritmo de k-means es utilizado en este trabajo para encontrar un conjunto de centroides que representen nuestras palabras visuales, las cuales son etiquetadas con un identificador en nuestro diccionario visual. Esta última etapa es ilustrada en el cuarto paso de la Figura 4.

Para representar las imágenes con el anterior diccionario visual, cada imagen pasa a través de la

malla, y cada parche de la imagen es reemplazado por las palabra visual más cercana en el diccionario visual (ver Figura 5). De esta forma, en un paso posterior cada imagen es representada por un histograma que cuenta la ocurrencia de las palabras visuales (de acuerdo al diccionario visual) en la imagen. En la siguiente sección, se muestra cómo utilizar el tan mencionado diccionario visual con el objetivo de construir los n-gramas de palabras visuales.

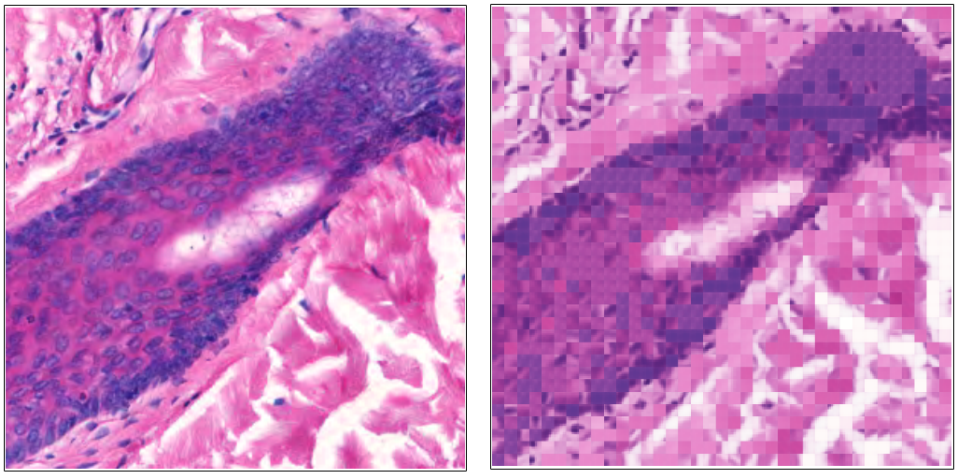

Figura 5. Ejemplo de una imagen representada usando el diccionario de palabras visuales. (a) Imagen original. (b) Representación por medio de palabras visuales.

7.2.2. Desarrollo de la bolsa de n-gramas de palabras visuales

En esta sección se presenta la segunda etapa para la construcción de los n-gramas visuales. Tal como ya se ha mencionado, en este paso se asume que ya existe un diccionario de palabras visuales para representar a las imágenes.

La idea intuitiva para la captura de las relaciones espaciales está inspirada por el uso de n-gramas en el área de clasificación de texto. En esta área los n-gramas de palabras son secuencias de n palabras consecutivas. Este tipo de secuencias simples ayudan a mantener ciertas relaciones entre las palabras. Debido a esto, la representación de bolsa de n-gramas puede tomar ventaja de conceptos como “Casa Blanca” el cual podría ser representado como un solo atributo en la representación vectorial. Sin embargo, la extracción de n-gramas visuales a partir de las imágenes no es tan simple como en el caso del texto. Mientras un documento de texto puede ser leído en una sola dirección,

las secuencias de los descriptores de imagen pueden ser obtenidos de muchas maneras diferentes (e.g., buscar secuencias horizontalmente, verticalmente, a un ángulo de θ grados, etc.) pues la dirección recae en lo que parece un plano 2D, esto debe ser considerado a la hora de la extracción de los n-gramas de las imágenes, ver Figura 6. Otra situación a considerar es cómo determinar la correcta dirección para interpretar los n-gramas. Un ejemplo sencillo para entender esto usando palabras son los 3-gramas compuestos por las mismas palabras pero diferente orden, esos 3-gramas normalmente tiene diferente significado. Por ejemplo, el 3-grama "oro de México" es altamente probable que se refiera al metal precioso de México, mientras el 3-grama "México de oro" seguramente se refiere a otra cosa. Por otro lado, en los n-gramas visuales que tienen la misma secuencia pero en diferente orientación (e.g., si una imagen fue rotada), por ejemplo el 3-grama 12-65-654 y 654-65-12 en la Figura 6, pudiera estar relacionado con el mismo patrón. En este experimento preliminar, se consideraron ambos patrones, 12-65-654 y 654-65-12, el mismo n-grama. De esta forma los n-gramas visuales generados son menos sensibles a los problemas de rotación.

123	**213**	**12**	**33**	65	34	43	673
254	**546**	**65**	**444**	346	637	546	456
45	**645**	**654**	**565**	456	456	54	45
34	43	673	123	213	12	33	43
637	546	456	254	546	65	444	546
456	54	45	45	645	654	565	54
34	43	673	123	213	12	33	33
637	546	456	254	546	65	444	444

Figura 6. El proceso para la construcción de n-gramas visuales a través del uso de una ventana. Para el recuadro sombreado (65) los n-gramas extraidos son: 65-12, 65-213, 65-546, 65-645, 65-654, 65-565, 65-444, 65-33.

Con el objetivo de construir los n-gramas visuales aplicamos el siguiente enfoque. Primeramente recordar que, por cada imagen se tiene un documento conteniendo la matriz de códigos de palabras visuales (ver la Figura 6). El Algoritmo 1 muestra el enfoque para obtener los n-gramas visuales. La idea principal es producir n-gramas ignorando la orientación en la que aparecen. Para construir n-gramas se itera sobre cada elemento $a_{i,j}$ de la matriz A (líneas 2 y 3) y se procede a extraer los

vecinos en línea recta (líneas 4 a 12). Esto es, se extraen secuencias utilizando los elementos entre $a_{i,j}$ y $a_{i+k,j+h}$, si y solo si éstos son parte de la línea recta que une a $a_{i,j}$ y $a_{i+k,j+h}$. Esto conduce a obtener n-gramas en dirección horizontal, vertical y diagonal. Las líneas de "si existe ..." (líneas 5 a 12) en el Algoritmo 1 son necesarias debido a que incluso para los elementos en las esquinas y bordes se trata de extraer todos los posibles vecinos (esta forma de presentarlo ayuda a que la explicación sea sencilla y fácil). Esta última parte nos deja con ocho posibles n-gramas para cada posición en la matriz. Finalmente, cada n-grama es normalizado para ser interpretado solamente de una forma y consecuentemente indexado como el mismo elemento en nuestro nuevo diccionario de n-gramas visuales (líneas 14 a 16).

Algoritmo 1 obtener n-gramas

Entrada: **A** (una matriz de $x \times y$ conteniendo los identificadores de las palabras visuales), **n** (la longitud de las secuencias visuales requeridas)

Salida: $L = (S_1, \ldots S_k); k = 1...l$ (una lista de las secuencias encontradas.)

1: $n = n - 1$
2: **Para** $i = 0$ **hasta** x **hacer**
3: **Para** $j = 0$ **hasta** y **hacer**
4: Crea una lista de secuencias temporales (LST) vacía
5: Si $(a_{i,j}, \ldots, a_{i-n,j})$ existe, añadir a LST
6: Si $(a_{i,j}, \ldots, a_{i-n,j-n})$ existe, añadir a LST
7: Si $(a_{i,j}, \ldots, a_{i,j-n})$ existe, añadir a LST
8: Si $(a_{i,j}, \ldots, a_{i+n,j-n})$ existe, añadir a LST
9: Si $(a_{i,j}, \ldots, a_{i+n,j})$ existe, añadir a LST
10: Si $(a_{i,j}, \ldots, a_{i+n,j+n})$ existe, añadir a LST
11: Si $(a_{i,j}, \ldots, a_{i,j+n})$ existe, añadir a LST
12: Si $(a_{i,j}, \ldots, a_{i+n,j+n})$ existe, añadir a LST
13: **Para** cada elemento secuencia E **en** LST **hacer**
14: **Si** $e_0 > e_n$ **entonces**
15: invertir(E)
16: **Fin Si**
17: **Fin Para**
18: Añadir elementos en LST a L
19: $j++$
20: **Fin Para**
21: $i++$
22: **Fin Para**

Una vez que tenemos el diccionario de n-gramas visuales se procede con la representación de la imagen. Para esto, cada imagen es representada por un histograma que contiene la ocurrencia de los n-gramas visuales encontrados en la imagen.

7.2.3. Clasificación de imágenes

Para llevar a cabo la clasificación de imágenes, éstas se representan a través de BoVN y se utilizan los histogramas como vectores de características para entrenar un clasificador. Para esto se utilizó una máquina de soporte vectorial (SVM, *Support Vector Machine*) por medio de la configuración estándar del algoritmo de optimización mínima secuencial de Weka (Hall y cols., 2009). Se utilizó un SVM debido a que, de entre otros enfoques éste ha mostrado ser muy efectivo utilizando la representación de BoVW (Boiman y cols., 2008). Además, el SVM ha sido utilizado en otros problemas similares de imágenes de histología con el objetivo de encontrar patrones visuales (Cruz-Roa, Caicedo, y González, 2011; Díaz y Romero, 2012).

7.3. Evaluación de la utilidad de los n-gramas de palabras visuales

Para la evaluación del enfoque propuesto, se construyó un clasificador binario para cada categoría. Para lograr esto, se tomaron como positivas las instancias pertenecientes a la categoría objetivo, y el resto como negativas (i.e., un enfoque estándar de uno contra el resto). Como se puede observar, cada problema binario está desbalanceado, en particular para las clases 3-7, lo que agrega dificultad al problema.

Se han llevado a cabo diversos experimentos para cada problema de clasificación. En esos experimentos, para extraer parches del conjunto de datos se han utilizado dos configuraciones: i) con parches de 8x8 pixeles (el cual denotamos como 8) y con parches de 16x16 pixeles (el cual denotamos como 16). Para cada experimento se ha realizado una validación cruzada de 10 pliegues. Es importante notar que, en nuestros experimentos de n-gramas, una configuración de orden n incluye a todos los n-gramas de menos o igual orden que n. La combinación de estas características es en esta forma, debido a que es una alternativa en la que los n-gramas han mostrado mejorar la clasificación de texto (Bekkerman y Allan, 2004; Tan y cols., 2002; S. Wang y Manning, 2009) (también se llevaron a cabo experimentos con representaciones separadas pero estos obtuvieron resultados inferiores, confirmando los resultados de algunos trabajos en tareas de categorización de texto). Para la extracción de los n-gramas visuales se cuenta con 400 palabras visuales (unigramas) y un diferente número de n-gramas para cada distinto valor de n (de 1 a 4). Esto último significa que, en un experimento de 3-gramas ($1 + 2 + 3grams$) se han combinado 400 unigramas más x bigramas más x bigramas más x trigramas para la BoVN. Además, es valioso mencionar que se

ha normalizado cada conjunto de atributos de forma individual (esto es que la suma del vector de palabras visuales de cada modalidad es uno). Finalmente, para los experimentos se han probado dos de los más populares esquemas de pesado de términos en categorización de texto. El primero es la frecuencia del término, que es denotado como TF. El pesado TF consiste en la utilización de un histogramas de valores de frecuencia en un vector de características, pero normalizado por la longitud del documento. Por otro lado, el esquema de pesado booleano construye vectores de características remplazando cada valor v del histograma por un 1 si v >0 y por un 0 en caso contrario.

En las siguientes secciones se explica el propósito, detalle y resultados para cada experimento, los cuales han sido elegidos cuidadosamente para analizar las diferentes propiedades de el uso de los n-gramas visuales para clasificar imágenes. El mejor resultado para cada serie de experimentos se denota con negritas. En las siguientes tablas reportamos el promedio (sobre todas las clases) de la medida F_1 y el promedio de el área bajo la curva ROC (AUC) de los siete problemas binarios en la colección de imágenes de histopatología BCC.

7.3.1. Comparación entre bolsa de palabras visuales y bolsa de n-gramas visuales

Esta serie de experimentos iniciales se ha realizado bajo cuatro diferentes escenarios con distinta configuración. Para el pesado de los términos tenemos: i) binario (BIN), y frecuencia de término (TF). Para el tamaño de los parches se tiene: i) 8x8 (8), y ii) 16x16 (16). Se muestra el promedio de la medida F_1 y el AUC de los siete problemas binarios para nuestra colección de imágenes de histología.

El primer experimento considera todas las palabras visuales contenidas en las imágenes. El objetivo de este experimento es determinar la efectividad de la clasificación de la tradicional BoVW bajo diferentes condiciones. Los experimentos de la Tabla 3 muestran que el parche de tamaño 8 y el pesado TF obtienen los mejores resultados. Lo cual de alguna forma era esperado dado que el parche de 8 está relacionado con un buen tamaño de resolución para cubrir la estructura biológica de las células (Cruz-Roa, Caicedo, y González, 2011). Por otro lado, también se piensa que el pesado TF en general es una buena opción dado que mantiene un conteo de los patrones visuales (un pesado binario intuitivamente solamente busca por la presencia o ausencia del atributo).

Una vez que tenemos una idea del rendimiento que puede ser obtenido bajo un enfoque estándar de BoVW, se analiza el rendimiento de la BoVN. Para lo anterior, primeramente se estudia el

Visual Words		
Config	*FM*	*AUC*
Bin-8	48.27	67.74
Bin-16	47.63	67.56
TF-8	**58.59**	**72.27**
TF-16	52.33	68.89

Tabla 3. Experimentos utilizando palabras visuales (Unigramas) a través de dos tipos de pesado de términos (TF y BIN) y dos distintos tamaños de parche (8 y 16).

rendimiento de BoVN a través de distintos números de n-gramas, esto es, se hace uso de los x más frecuentes n-gramas (para este experimento se utilizan los bigramas) para construir nuestro diccionario visual. Esta reducción es necesaria dado que el número inicial de bigramas es de varias decenas de miles. El siguiente experimento presenta un estudio para conocer como afecta el número de características que tomamos. La Tabla 4 muestra los resultados obtenidos con la mejor configuración (parches de 8 y pesado TF) variando el número de los bigramas más frecuentes. Para estos experimentos podemos observar que utilizar 2500 bigramas resulta en el mejor rendimiento, sólo un poco mejor que el experimento que utiliza 5000. Además, se puede observar que existe una diferencia de al menos 6 % en el 64.31 % de la columna 2.5K de la medida F_1 promediada obtenida por los unigramas-bigramas, contra el 58.59 % de la medida F_1 promediada en los experimentos utilizando solamente unigramas (lo cual es una tradicional BoVW). Se tiene la idea de que esto es debido a que los pares de palabras visuales logran capturar buenos patrones visuales, lo cual de alguna forma está en armonía con la evidencia en categorización de texto. Dado este escenario, se tomará ventaja de esta representación en los siguientes experimentos.

Frequency threshold					
Config	*1.5K*	*2.5K*	*5K*	*7.5K*	*10K*
1+2grams	63.73	**64.31**	64.03	62.24	61.63

Tabla 4. Experimentos utilizando Bigramas para analizar el impacto de la dimensionalidad.

Los experimentos en la Tabla 5 utilizan bigramas visuales con el fin de examinar su comportamiento bajo las mismas condiciones que las palabras visuales. De la Tabla 5 se puede observar que el pesado TF con parches de tamaño 8, de nuevo supera a las otras configuraciones. También es valioso notar que, de manera general, y bajo las mismas condiciones la combinación de atributos

de bigramas de palabras visuales son mejores que utilizar solamente las palabras visuales.

Unigrams vs Uni+Bigrams				
Config	*F-Measure*		*AUC*	
	1grams	*1+2grams*	*1grams*	*1+2grams*
Bin-8	48.27	59.50	67.74	72.54
Bin-16	47.63	56.67	67.56	70.46
TF-8	58.9	**64.31**	72.27	**76.03**
TF-16	52.33	56.09	68.89	71.17

Tabla 5. Experimentos utilizando secuencias de palabras visuales (Uni-Bi-gramas) a través de dos tipos de pesado de términos (TF y BIN) y dos distintos tamaños de parche (8 y 16) .

El último experimento de esta sección se muestra en la Tabla 6. Esta tabla presenta los resultados de los experimentos en un enfoque de BoVN con distintos valores de n. Se llevaron a cabo experimentos para determinar si considerar n-gramas de más alto orden que 2, podría mejorar el rendimiento de el clasificador. De los resultados en la Tabla 6 se puede notar que la mejor configuración se mantiene en $1 + 2$gramas. Suponemos que esto es debido a las siguientes razones. La primera está relacionada con el tamaño de las secuencias: es bien sabido que entre más grande la n para los n-gramas, mayor es el número de instancias que son requeridas para encontrar esas secuencias de tamaño n (Tan y cols., 2002). La segunda está relacionada con la alta dimensionalidad: utilizar secuencias más largas produce grandes vocabularios, los cuales generan vectores de características dispersos.

Experiments with n-grams	
Config	*FM*
1grams	58.59
1+2grams	**64.31**
1+2+3grams	62.69
1+2+3+4grams	61.34

Tabla 6. Experimentos utilizando secuencias de palabras visuales (de Unigramas a Tetragramas) para el análisis del impacto en la longitud de la secuencia.

7.3.2. Análisis detallado por clase para unigramas y bigramas de palabras visuales

En esta sección se presentan los resultados por clase. Los experimentos de esta sección consideran la mejor configuración de la Tabla 3 utilizando unigramas visuales contra la mejor configuración de la Tabla 5 utilizando bigramas visuales.

En la Tabla 7 y 8 mostramos respectivamente la medida F_1 y la AUC obtenida para la clase positiva en cada categoría. Para estos experimentos, se ha llevado a cabo una validación cruzada de 10 pliegues utilizando unigramas y bigramas en cada uno de los siete problemas binarios. También en la columna "(b-a) ganancia/pérdida" se muestra la ganancia o pérdida (en la medida F_1 o AUC) causada por el uso de bigramas.

Detailed F-Measure by class			
Class	*(a) 1grams*	*(b) 1+2grams*	*(b-a) gain/loss*
1	86.10	90.70	4.6
2	94.80	95.50	0.7
3	74.40	83.40	9.0
4	36.80	50.80	14.0
5	35.80	52.50	16.7
6	48.00	43.60	-4.4
7	34.20	33.70	-0.5

Tabla 7. Experimentos detallados por clase para el F_1 utilizando palabras visuales (Unigramas) contra secuencias de palabras visuales (Uni-Bi-gramas).

Los experimentos en las Tablas 7 y 8 muestran que a través del uso de los bigramas visuales, mayores beneficios (cuarta columna) son obtenidos en comparación con los unigramas, para las clases 1,3,4 y 5, las cuales son respectivamente células basales carcinoma, epidermis, folículo piloso y glándulas ecrinas. De estas clases, la más importante es la primera, esto es debido a que es la única relacionada con el diagnóstico de cáncer. Se piensa que los buenos resultados en la clase 1 fueron obtenidos debido a que la mayoría de esas imágenes son caracterizadas por estructuras de células de tumor que tienen núcleos grandes y obscuros, los cuales son capturados por los bigramas visuales. Por otro lado, las palabras visuales parecen ser muy competitivas o mejores en las clases 2, 6 y 7, las cuales son colágeno, glándulas sebaceas y infiltrado inflamatorio (ninguna de ellas relacionada con el diagnóstico de cáncer). Estas son imágenes en dónde las relaciones es-

Detailed AUC by class			
Class	*(a) 1grams*	*(b) 1+2grams*	*(b-a) gain/loss*
1	89.00	92.50	3.5
2	76.40	79.20	2.8
3	84.00	90.90	6.9
4	62.60	68.80	6.2
5	62.80	71.60	8.8
6	68.70	66.90	-1.8
7	62.40	62.30	-0.1

Tabla 8. Experimentos detallados por clase para el AUC utilizando palabras visuales (Unigramas) contra secuencias de palabras visuales (Uni-Bi-gramas).

paciales estructuradas de sus componentes, no son suficientemente abundantes para ser capturadas por el uso de bigramas principalmente por: i) ser demasiado simples (e.g., colageno e infiltrado inflamatorio) o ii) ser altamente complejas con demasiada variabilidad visual (glándulas sebaceas). Suponemos que para obtener mejores elementos visuales, estos problemas necesitaran: i) técnicas de NLP que vayan más allás de encontrar relaciones espaciales locales, y ii) un estudio más adecuado de los parámetros utilizados (e.g., tamaño del parche, secuencias largas de palabras visuales, o descriptores alternativos).

7.3.3. Comparación contra otros enfoques típicos

Además del SVM utilizando palabras visuales y n-gramas visuales como atributos, también mostramos experimentos utilizando otro enfoque que se ha convertido en un clásico en las palabras visuales; el clasificador basado en modelos del lenguaje. Tal cómo se ha explicado en la Sección 2, los modelos del lenguaje han sido utilizados en trabajos previos para construir clasificadores (Tirilly y cols., 2008; Wu y cols., 2007). En esta propuesta, se ha reproducido un clasificador basado en modelos del lenguaje similar al utilizado en (Tirilly y cols., 2008), el cual está basado en el *CMU-Cambridge Statistical Language Modeling Toolkit v2* (Clarkson y Rosenfeld, 1997). El objetivo de este experimento es comparar nuestra propuesta con metodologías alternativas en el estado del arte que también usen secuencias de palabras visuales. El clasificador basado en modelos del lenguaje utiliza $1 + 2 + 3$gramas (configuraciones de 2-gramas hasta 10-gramas fueron probadas)

el resto de los parámetros del software para construir modelos de lenguaje específicos fueron dejados por omisión (e.g., suavizado de *good turing dicount* y el *backoff*). El clasificador que hemos programado trabaja de la siguiente manera:

- Para cada problema binario, se toman los documentos de entrenamiento y se construyen dos modelos del lenguaje (uno para la clase positiva y uno para la negativa).
- Para cada documento de prueba, se mide la distancia (utilizando la regla de la cadena con probabilidad y la perplejidad) contra el modelo positivo y negativo, posteriormente se asigna la clase a la categoría más cercana.

La Tabla 9 muestra los resultados de los experimentos comparando el clasificador de modelos del lenguaje (CML) y nuestro SVM-BoVN. Para este experimento, se puede observar que al menos para este problema y bajo las mismas condiciones el CML no proporciona un mejor rendimiento que el SVM-BoVN. Pensamos que esto es por el problema de alto desbalance en las clases y la naturaleza probabilista de los modelos del lenguaje, en un conjunto de imágenes dónde se tienen muy pocos documentos para la construcción de modelos de lenguaje precisos en algunas clases positivas.

LMC vs Uni+Bigrams				
Config	F-Measure		AUC	
	LMC	1+2grams	LMC	1+2grams
TF-8	53.0	**64.31**	69.89	**76.03**
TF-16	48.31	56.09	72.21	71.17

Tabla 9. Experimentos utilizando secuencias de palabras visuales (Uni-Bi-Gramas) comparadas con CML.

7.4. Conclusiones de los resultados preliminares

En este experimento preliminar se propone una extensión a la representación estándar BoVW. La extensión propuesta se centra en la extracción de patrones secuenciales y los utiliza como atributos para inducir un modelo de clasificación. Es valioso notar que, aunque la idea general de n-gramas en palabras visuales ha sido explorada para trabajos de recuperación de información, modelos del

lenguaje, y selección de características, hasta donde sabemos, esta no ha sido utilizada bajo el esquema en el que lo utilizamos, ni como atributos para un algoritmo de aprendizaje automático en la tarea de clasificación de imágenes. De esta forma, como una primera aplicación se han probado las ideas en una colección de imágenes de histopatología. Con el objetivo de obtener las secuencias de patrones visuales, se extraen n-gramas que de alguna forma sean análogos a las ideas en NLP, pero tomando en cuenta las particularidades del dominio de imágenes. Los resultados experimentales bajo diferentes condiciones y configuraciones han mostrado que el uso de los n-gramas visuales como atributos son útiles, debido a que estos permiten mejorar la efectividad en la clasificación contra el tradicional enfoque de BoVW y el CML. Pensamos que esto es debido a que el método logra encontrar patrones visuales, los cuales pueden ser difíciles de obtener a partir de las palabras visuales individuales.

8. Conclusiones

En este documento se describe parte del trabajo que se ha realizado en el periodo de Enero-Diciembre 2013, y parte del trabajo que se planea llevar a cabo durante el programa de Doctorado. El principal interés de esta investigación se centra en la intersección de las áreas de HLCV y NLP. De esta forma, el trabajo se enfocará en hacer clasificación y recuperación de imágenes mejorando los métodos basados en palabras visuales mediante la introducción de representaciones visuales que pueden ser análogos a representaciones textuales. A través de la utilización de métodos de NLP se pretende capturar otro tipo de información, que actualmente no es considerada por la BoVW estándar. Por ejemplo, el uso de la información contextual y de alto nivel que existe entre los elementos de una imagen, y que no es considerada en la BoVW tradicional. Dado que la utilización de la información contextual es un factor común en muchas tareas de NLP, como una idea inicial, se propone el uso de lo que sería la extensión natural de la BoVW; la utilización de n-gramas de palabras visuales como atributos para un clasificador. En este sentido, se presentó como alternativa la Bolsa de n-gramas de palabras visuales (BoNVW). Actualmente, se trabaja en la extracción de secuencias frecuentes maximales que podrían ser usadas como atributos o ser utilizadas como la base de un clasificador basado en reglas de asociación. Asimismo, se está desarrollando un método para fusionar los atributos de palabras visuales y los n-gramas de palabras visuales de mejor forma. Para ello, se está comenzando a trabajar con métodos de fusión tardía, fusión temprana, que logren

combinar de mejor manera la información de cada espacio de atributos.

9. Publicaciones

Los resultados preliminares contenidos en esta propuesta de investigación, se encuentran publicados en:

- ⋆ *Bag-of-visual-ngrams for histopathology image classification*, A. Pastor López-Monroy, Manuel Montes-y-Gómez, Hugo Jair Escalante, Ángel Cruz-Roa and Fabio A. González, Proc. SPIE 8922, IX International Seminar on Medical Information Processing and Analysis, 89220P (November 19, 2013).

Referencias

Argamon, S., Koppel, M., Pennebaker, J. W., y Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM*, *52*(2), 119–123.

Avni, U., Greenspan, H., Konen, E., Sharon, M., y Goldberger, J. (2011). X-ray categorization and retrieval on the organ and pathology level, using patch-based visual words. *IEEE Transactions on Medical Imaging*, *30*(3), 733–746.

Bekkerman, R., y Allan, J. (2004). *Using bigrams in text categorization* (Inf. Téc.). Department of Computer Science, University of Massachusetts, Amherst.

Blei, D. M., Ng, A. Y., y Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*, 993–1022.

Boiman, O., Shechtman, E., y Irani, M. (2008). In defense of nearest-neighbor based image classification. En *Ieee computer society conference on computer vision and pattern recognition* (pp. 1–8).

Cao, L., y Fei-Fei, L. (2007). Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. En *Ieee 11th international conference on computer vision* (pp. 1–8).

Clarkson, P., y Rosenfeld, R. (1997). Statistical language modeling using the cmu-cambridge toolkit. En *Proceedings of eurospeech* (Vol. 97, pp. 2707–2710).

Cruz-Roa, A., Caicedo, J. C., y González, F. A. (2011). Visual pattern mining in histology image collections using bag of features. *Artificial intelligence in medicine*, *52*, 91–106.

Cruz-Roa, A., Díaz, G., Romero, E., y González, F. A. (2011). Automatic annotation of histopathological images using a latent topic model based on non-negative matrix factorization. *Journal of Pathology Informatics*, *4*.

Csurka, G., Dance, C. R., Fan, L., Willamowski, J., y Bray, C. (2004). Visual categorization with bags of keypoints. En *International workshop on statistical learning in computer vision* (Vol. 1, p. 22).

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., y Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407.

Deselaers, T., Pimenidis, L., y Ney, H. (2008). Bag-of-visual-words models for adult image classification and filtering. En *19th international conference on pattern recognition* (pp. 1–4).

Díaz, G., y Romero, E. (2012). Micro-structural tissue analysis for automatic histopathological image annotation. *Microscopy Research and Technique*, *75*, 343–358.

Escalante, H. J., Solorio, T., y Gómez, M. Montes-y. (2011). Local histograms of character n-grams for authorship attribution. En *Proceedings of the 49th annual meeting of the association for computational linguistics* (pp. 288–298).

Galleguillos, C., y Belongie, S. (2010). Context based object categorization: A critical survey. *Computer Vision and Image Understanding*, *114*, 712–722.

García-Hernández, R. A., Martínez-Trinidad, J. F., y Carrasco-Ochoa, J. A. (2004). A fast algorithm to find all the maximal frequent sequences in a text. En *Progress in pattern recognition, image analysis and applications* (pp. 478–486). Springer.

García-Hernández, R. A., Martínez-Trinidad, J. F., y Carrasco-Ochoa, J. A. (2006). A new algorithm for fast discovery of maximal sequential patterns in a document collection. En *Computational linguistics and intelligent text processing* (pp. 514–523). Springer.

Giannakopoulos, G., Karkaletsis, V., Vouros, G., y Stamatopoulos, P. (2008). Summarization system evaluation revisited: N-gram graphs. *ACM Transactions on Speech and Language Processing*, *5*(3), 5.

Giannakopoulos, G., Mavridi, P., Paliouras, G., Papadakis, G., y Tserpes, K. (2012). Representation

models for text classification: a comparative analysis over three web document types. En *Proceedings of the 2nd international conference on web intelligence, mining and semantics* (p. 13).

Gönen, M., y Alpaydın, E. (2011). Multiple kernel learning algorithms. *Journal of Machine Learning Research*, *12*, 2211–2268.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., y Witten, I. (2009). The weka data mining software: An update. *SIGKDD Explorations*, *11*.

Jamieson, M., Fazly, A., Dickinson, S., Stevenson, S., y Wachsmuth, S. (2007). Learning structured appearance models from captioned images of cluttered scenes. En *Ieee 11th international conference in computer vision* (pp. 1–8).

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. En *Proceedings of the 10th european conference on machine learning* (pp. 137–142).

Koppel, M., Argamon, S., y Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, *17*(4), 401–412.

Koppel, M., y Schler, J. (2003). Exploiting stylistic idiosyncrasies for authorship attribution. En *Workshop on computational approaches to style analysis and synthesis* (pp. 69–72).

Kuncheva, L. (2005). Combining pattern classifiers. *Wiley Press, New York*, 241–259.

Lavelli, A., Sebastiani, F., y Zanoli, R. (2004). Distributional term representations: an experimental comparison. En *Proceedings of the thirteenth acm international conference on information and knowledge management* (pp. 615–624).

Lazebnik, S., Schmid, C., y Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. En *Ieee computer society conference on computer vision and pattern recognition, 2006* (Vol. 2, pp. 2169–2178).

Lebanon, G., Mao, Y., y Dillon, J. V. (2007). The locally weighted bag of words framework for document representation. *Journal of Machine Learning Research*, *8*(10), 2405–2441.

Li, Z., Xiong, Z., Zhang, Y., Liu, C., y Li, K. (2011). Fast text categorization using concise semantic analysis. *Pattern Recognition Letters*, *32*(3), 441–448.

López-Monroy, A. P., Montes-y-Gómez, M., Escalante, H. J., , y Villatoro-Tello, E. (2013). Inaoe's participation at pan'13: Author profiling task. En *Notebook of clef-pan 2013*.

López-Monroy, A. P., Montes-y-Gómez, M., Villaseñor-Pineda, L., Carrasco-Ochoa, J. A., y Martínez-Trinidad, J. F. (2012). A new document author representation for authorship attribution. En *Mexican conference in pattern recognition* (pp. 283–292). Springer.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, *60*(2), 91–110.

McCallum, A., y Nigam, K. (1998). A comparison of event models for naive bayes text classification. En *Aaai-98 workshop on learning for text categorization* (Vol. 752, pp. 41–48).

Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal*, *26*(4), 354–359.

Nowak, E., Jurie, F., y Triggs, B. (2006). Sampling strategies for bag-of-features image classification. En *Computer vision–eccv 2006* (pp. 490–503). Springer.

Quack, T., Ferrari, V., Leibe, B., y Van Gool, L. (2007). Efficient mining of frequent and distinctive feature configurations. En *Ieee 11th international conference on computer vision.* (pp. 1–8).

Schler, J., Koppel, M., y Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science*, *60*, 9–26.

Schler, J., Koppel, M., Argamon, S., y Pennebaker, J. (2006). Effects of age and gender on blogging. En *Proceedings of 2006 aaai spring symposium on computational approaches for analyzing weblogs* (pp. 199–205).

Scott, M. (2001). Comparing corpora and identifying key words, collocations, and frequency distributions through the wordsmith tools suite of computer programs. *Small corpus studies and ELT*, 47–67.

Sivic, J., Russell, B. C., Efros, A. A., Zisserman, A., y Freeman, W. T. (2005). *Discovering object categories in image collections* (Inf. Téc.). CSAIL, MIT.

Sivic, J., y Zisserman., A. (2003). Video google: A text retrieval approach to object matching in videos. En *Proceedings of the international conference on computer vision.*

Sivic, J., y Zisserman, A. (2004). Video data mining using configurations of viewpoint invariant regions. En *Proceedings of the 2004 ieee computer society conference on computer vision and pattern recognition* (Vol. 1, pp. I–488).

Stamatatos, E. (2009). A survey on modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, *60*, 538–556.

Tan, C. M., Wang, Y. F., y Lee, C. D. (2002). The use of bigrams to enhance text categorization. *Information processing and management*, *38*, 529–546.

Tirilly, P., Claveau, V., y Gros, P. (2008). Language modeling for bag-of-visual words image categorization. En *Acm proceedings of the 2008 international conference on content-based image and video retrieval* (pp. 249–258).

Tirilly, P., Claveau, V., y Gros, P. (2009a). *A review of weighting schemes for bag of visual words image retrieval* (Inf. Téc.). Technical report, TEXMEX - INRIA - IRISA.

Tirilly, P., Claveau, V., y Gros, P. (2009b). *A review of weighting schemes for bag of visual words image retrieval* (Inf. Téc.). IRISA, Rennes, France.

Turney, P., y P., P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188.

Wang, H., Ullah, M. M., Klaser, A., y Laptev, I. (2009). Evaluation of local spatio-temporal features for action recognition. En *Proceedings of the british machine vision conference* (pp. 1–11).

Wang, J., Li, Y., Zhang, Y., Wang, C., Xie, H., Chen, G., y cols. (2011). Bag-of-features based medical image retrieval via multiple assignment and visual words weighting. *IEEE transactions on medical imaging*, *30*(11), 1996–2011.

Wang, S., y Manning, C. D. (2009). Baselines and bigrams: Simple, good sentiment and topic classification. En *Proceedings of the 50th annual meeting of the association for computational linguistics* (pp. 90–94).

Wang, X., y Grimson, E. (2007). Spatial latent dirichlet allocation. En *Advances in neural information processing systems* (pp. 1577–1584).

Winn, J., Criminisi, A., y Minka, T. (2005). Object categorization by learned universal visual dictionary. En *Tenth ieee international conference on computer vision* (Vol. 2, pp. 1800–1807).

Wu, L., Li, M., Li, Z., Ma, W. Y., y Yu, N. (2007). Visual language modeling for image classification. En *Acm proceedings of the international workshop on workshop on multimedia information retrieval* (pp. 115–124).

Yuan, J., Wu, Y., y Yang, M. (2007). Discovery of collocation patterns: from visual words to visual phrases. En *Ieee in computer vision and pattern recognition* (pp. 1–8).

Yuan, J., Yang, M., y Wu, Y. (2011). Mining discriminative co-occurrence patterns for visual

recognition. En *Ieee conference on computer vision and pattern recognition* (pp. 2777–2784).

Zhang, J., Marszalek, M., Lazebnik, S., y Schmid, C. (2007). Local features and kernels for classification of texture and object categories: a comprehensive study. *International Journal of Computer Vision*, *73*, 213–238.

Zheng, Q. F., Wang, W., y Gao, W. (2006). Effective and efficient object-based image retrieval using visual phrases. En *Acm proceedings of the 14th annual acm international conference on multimedia* (pp. 77–80).

Zhixing, L., Zhongyang, X., Yufang, Z., Chunyong, L., y Kuan, L. (2010). Fast text categorization using concise semantic analysis. *Pattern Recognition Letters*, *32*, 441–448.