Dpto. de Teoría de la Señal y Comunicaciones
Universidad Carlos iii de Madrid

TESIS DOCTORAL

GENERATIVE MODELS FOR IMAGE SEGMENTATION AND REPRESENTATION

Autor: Iván González Díaz
Director: Dr. Fernando Díaz De María

Leganés, 2011

Tesis Doctoral:
GENERATIVE MODELS FOR IMAGE SEGMENTATION AND REPRESENTATION

Autor:
IVÁN GONZÁLEZ DÍAZ

Director:
DR. FERNANDO DÍAZ DE MARÍA

El tribunal nombrado para juzgar la tesis doctoral arriba citada, compuesto por los doctores

Presidente:

Vocales:

Secretario:

acuerda otorgarle la calificación de

Leganés, a

RESUMEN EXTENDIDO

En este resumen extendido se presenta una descripción de los aspectos más relevantes de la presente Tesis doctoral. En particular, se describe el área de interés y plantea la motivación del trabajo realizado. Posteriormente, se explicarán las contribuciones originales más relevantes para, finalmente, establecer las conclusiones más significativas y enumerar una serie de líneas de investigación que surgen a partir del trabajo realizado.

Motivación de la Tesis

En un área de investigación tan compleja como la visión artificial, esta Tesis se centra en dos campos bien definidos: segmentación y reconocimiento.

La segmentación es la tarea de buscar grupos de píxeles relacionados (regiones) en una imagen y, aunque constituye uno de los problemas más antiguos y estudiados en visión artificial, aún no puede considerarse cerrado. En estadística, este problema se conoce habitualmente como *clustering* o agrupamiento y se trata de un campo muy estudiado, en el que se han propuesto cientos de diferentes algoritmos [Jain and Dubes, 1988], [Jain et al., 2004]. En esta Tesis, la tarea de segmentación se ha abordado desde dos puntos de vista diferentes: a) segmentación no supervisada en secuencias audiovisuales y b) segmentación supervisada con etiquetado semántico.

La segmentación no supervisada divide una imagen en un conjunto de segmentos que son homogéneos con respecto a una o varias medidas de homogeneidad. Por tanto, estos algoritmos se relacionan con las primeras etapas del sistema visual humano.

En la primera parte de la Tesis, se propondrán algoritmos para la segmentación no supervisada de imágenes en secuencias audiovisuales, escenario en el cual se dispondrá de dos tipos de información: información estática relativa a propiedades de la imagen tales como el color o la textura, e información dinámica asociada al movimiento existente en la escena.

En contraposición, la segmentación con etiquetado semántico emplea información previa y aprendida sobre los objetos a segmentar, de modo que el objetivo de dichos algoritmos no es únicamente proporcionar un conjunto de regiones sino también el etiquetado o categoría semántica (avión, cielo, coche, carretera, etc.) a la que pertenece cada píxel de una imagen. Por lo tanto, este campo de la visión artificial está fuertemente interrelacionado con otras tareas como el reconocimiento y la detección de objetos, siendo necesario un entrenamiento previo de los algoritmos a partir de muestras etiquetadas.

Por otro lado, el problema de reconocimiento en imágenes supone un paso más allá de la simple detección de instancias, pues propone la caracterización de conceptos semánticos genéricos. Es, por tanto, necesaria una capacidad de generalización a partir de distintas realizaciones de una misma categoría (por ejemplo, diferentes instancias del concepto coche pueden ser muy diferentes en función de aspectos tales como la escala, el punto de vista o variaciones propias del objeto: color, forma, etc.). En esta Tesis, el reconocimiento de imágenes se ha enfocado como un problema de detección binario, de modo que una misma imagen puede contener varias categorías semánticas.

La segunda parte de la Tesis explorará el problema de la representación de imágenes, el cual proporcionará un marco para abordar las tareas particulares del reconocimiento, detección y segmentación en categorías semánticas.

Todos los algoritmos desarrollados en la Tesis hacen uso de modelos probabilísticos generativos, los cuales tratan de modelar la función de densidad de probabilidad conjunta de los datos y sus etiquetas. Partiendo de soluciones estadísticas de carácter general bien conocidas en la literatura, se han desarrollado extensiones y modelos avanzados para el campo de la visión artificial.

Metodología y aportaciones originales de la Tesis

Como se ha comentado, esta Tesis puede dividirse en dos partes: en la primera se han desarrollado algoritmos de segmentación espacio-temporal de imágenes en secuencias audiovisuales, mientras que en la segunda, se han diseñado modelos probabilísticos para el reconocimiento y segmentación de objetos en imágenes.

Con respecto a la segmentación espacio-temporal, el escenario de trabajo es el siguiente: una vez que un vídeo se ha dividido en varias tomas (entiéndase como toma un grupo de planos capturados sin cortes en la grabación, y que representan una acción continua en tiempo o espacio), varios planos consecutivos (necesarios para la estimación de movimiento) son extraídos en cada toma con el objetivo de proporcionar una segmentación sobre un plano particular (considerado el plano clave o *keyframe*). En este proceso se persiguen varios objetivos, los cuales constituyen nuestras contribuciones originales:

- Partiendo de una solución probabilística bien conocida como son los Modelos de Mezclas [Titterington et al., 1985], se ha propuesto un algoritmo que fusiona dos fuentes de información obtenidas de las secuencias de video: espacial (color, localización) y temporal (movimiento). La incorporación de información de movimiento ha permitido obtener regiones con mayor significado semántico, pues regiones “a priori” heterogéneas pueden ser combinadas si exhiben un patrón común de movimiento.

- Para la estimación de movimiento se han empleado técnicas robustas de *block-matching* que proporcionan una adaptación a diferentes tipos de movimiento: magnitud y patrones, escenas con fondos complejos, variaciones de escala, etc.

- Se ha diseñado un algoritmo jerárquico que permite implementar una estrategia de división (*splitting*) sobre las regiones. De este modo, uno de los parámetros libres en los modelos de mezclas, como es el número de componentes, es automáticamente calculado. Además, la utilización de distribuciones “a priori” ha permitido manejar tanto las nuevas componentes (nuevas regiones en la segmentación) como aquellas que no han sido modificadas.

- Por último, el proceso de división de regiones es controlado mediante un módulo de decisión que incorpora características de medio-nivel (*mid-level features*) espacio-temporales. Dichas características modelan propiedades encontradas en los conceptos u objetos del mundo real tales como regularidad, adyacencia o patrones habituales de movimiento (traslación, rotación, etc.).

Los algoritmos desarrollados en la primera parte de la Tesis han sido evaluados en una base de datos orientada a la extracción de información multimedia como es la base de datos de noticias de TRECVID 2006 [National Institute of Standards and Technology, 2006]. En nuestros experimentos se ha demostrado cómo las diferentes contribuciones permiten obtener segmentaciones más cercanas a ejemplos anotados por humanos, tanto desde un punto de vista general como utilizando evaluaciones orientadas al objeto u objetos de interés en la escena.

En la segunda parte de la Tesis se ha estudiado la aplicación de Modelos de Tópicos Latentes a la representación de imágenes. Estos modelos se basan en un proceso generativo que representa las imágenes como mezclas de una serie de tópicos de valor semántico, los cuales dan lugar a características visuales bien definidas. Así, es de esperar que una imagen se pueda expresar como una mezcla de tópicos, por ejemplo, avión o cielo. A su vez, cada tópico en particular dará lugar a descriptores visuales en ciertas áreas de la imagen. Los dos ejemplos más característicos encontrados en la literatura son *Probabilistic Latent Semantic Analysis* o PLSA [Hofmann, 2001], y *Latent-Dirichlet Allocation* o LDA [Blei et al., 2003]. Ambos métodos, concebidos en su origen para analizar textos, se apoyan en el paradigma de Bolsa de Palabras (*Bag of Words*) [Sivic and Zisserman, 2003, Csurka et al., 2004], según el cual conocer el orden de las palabras dentro de un documento no es necesario para lograr una correcta catalogación. Esta hipótesis que no es del todo cierta para documentos textuales, lo es aún menos en imágenes, donde el contenido visual se organiza siguiendo una estructura espacial muy bien definida.

Por lo tanto, en esta Tesis se han extendido dichos modelos para lograr una adaptación satisfactoria al problema de representación de imágenes que nos permita obtener información relevante tal como la aparición de los tópicos en las imágenes (clasificación o reconocimiento) o incluso la detección del área en la que aparece cada tópico (segmentación). En particular, se han propuesto dos modelos *Region-Based Latent Topic Model* o RBLTM y *Region-Based Latent Dirichlet Allocation* o RBLDA, en los que destacan las siguientes contribuciones:

- Extienden los modelos básicos PLSA y LDA para incorporar elementos que modelen la distribución espacial de los tópicos en una imagen.

- Incorporan segmentaciones no supervisadas como las desarrolladas en la primera parte de la Tesis y modelos cooperativos que permiten el intercambio de información entre las regiones de una imagen. Este proceso permite generar nuevas segmentaciones de valor semántico en las que tópicos o conceptos relacionados (por ejemplo cielo/avión) tienden a aparecer espacialmente conectados.

- Mejoran los modelos básicos de apariencia, aquellos implementados mediante modelos factorizables y distribuciones multinomiales, con nuevas propuestas que permiten capturar las relaciones no lineales entre descriptores locales que pertenecen a la misma región.

- Proponen marcos de trabajo flexibles para escenarios no supervisados, parcialmente y totalmente supervisados. Además, admiten dos tipos de anotaciones: etiquetas globales a nivel de imagen y etiquetado a nivel de región, bien mediante la utilización de *bounding-boxes*, o bien mediante el empleo de segmentaciones a nivel pixelar.

- Utilizan algoritmos de entrenamiento basados en inferencia variacional [Jordan et al., 1999], los cuales permiten resolver modelos gráficos muy complejos a través de aproximaciones más sencillas. Así, en cada caso un modelo variacional simplificado trata de aproximarse al original de forma que la divergencia de Kullback-Leibler [Kullback and Leibler, 1951] entre ambos se minimice.

De nuevo, los modelos desarrollados se han evaluado en bases de datos estándares utilizadas en visión artificial. En particular, las bases de datos empleadas pertenecen a las tareas de clasificación y segmentación de imágenes de la competición PASCAL VOC 2010 [Everingham et al., 2010]. Los modelos generados han sido comparados con varios algoritmos tomados como referencia, así como con otras alternativas encontradas en la literatura de Modelos de Tópicos Latentes. Por último, cabe destacar la inclusión de una comparativa de nuestra mejor propuesta frente a los resultados oficiales de la competición.

Conclusiones

A lo largo de la Tesis se han propuesto un serie de modelos generativos para solventar dos problemas que, si bien son muy conocidos, aún permanecen abiertos en visión artificial: la segmentación y la representación de imágenes.

Como se ha mencionado, la segmentación se ha planteado desde dos puntos de vista diferentes:

En el caso de la segmentación no supervisada se ha propuesto un algoritmo que trabaja sobre secuencias audiovisuales, estima el movimiento, y fusiona información temporal y espacial (color, coordenadas espaciales y vectores de movimiento) para dar lugar a la segmentación de un plano clave o *keyframe*. El algoritmo propuesto se basa en Modelos de Mezcla de Gaussianas a los que se incorporan distribuciones "a priori" sobre los parámetros con el fin de hacerlo adaptativo. La solución adaptativa permite, bajo un esquema jerárquico, proponer una solución iterativa de división (*splitting*) mediante la cual nuevas regiones se van añadiendo en cada iteración. Así, se ha logrado ajustar de modo automático el número de componentes de la mezcla, parámetro inherentemente libre en este tipo de propuestas.

Además, se ha desarrollado un módulo de decisión que utiliza descriptores espacio-temporales de medio-nivel (*mid-level features*), el cual ha resultado clave para considerar la inclusión o no de nuevas regiones en cada caso. Dado que las características de medio nivel modelan propiedades encontradas en los objetos/conceptos del mundo real, su empleo ha permitido lograr segmentaciones más cercanas a las que haría un ser humano.

Las pruebas realizadas sobre la base de datos TRECVID 2006 han mostrado que el algoritmo propuesto mejora los resultados de otras técnicas encontradas en el estado del arte. Además, a tenor de los resultados, podemos atribuir una influencia notable al módulo de decisión que emplea las características de medio nivel que modelan propiedades de los objetos y los patrones de movimiento en el mundo real. Sin embargo, la evaluación objetiva arroja ciertas dudas, pues los resultados numéricos no pueden considerarse concluyentes. Desde nuestro punto de vista esto es debido, no tanto a los propios resultados del algoritmo, como al hecho de que las medidas de calidad empleadas en la evaluación no se ajustan siempre a la percepción que los humanos tienen de una escena.

Los algoritmos de segmentación no supervisada constituyen una de las entradas a nuestras propuestas en la segunda parte de la Tesis, enfocadas al análisis y representación de imágenes. En este caso, tomando como base los modelos de tópicos latentes, se han propuesto extensiones de los mismos que permiten modelar la localización de los conceptos en las imágenes. El primero de los algoritmos propuestos, RBLTM, propone un modelo cooperativo en el que las regiones de una imagen interactúan para dar lugar a una representación espacialmente coherente de la misma. Nuestros experimentos han demostrado cómo el RBLTM mejora al algoritmo básico sobre el que se ha implementado, PLSA.

Sin embargo, ciertas limitaciones detectadas en el RBLTM han impulsado el diseño de otro modelo más avanzado: RBLDA. Éste, basado ahora en LDA, introduce ciertas soluciones para las debilidades del primero, pudiendo destacar aspectos tales como: la introducción de modelos sobre la distribución de los tópicos a nivel de un *corpus* de imágenes (en PLSA y RBLTM únicamente se incluyen modelos a nivel de cada imagen), modelos avanzados de apariencia que permiten exploter las relaciones entre descriptores que pertenecen a una misma región, un modelo de contexto nuevo que ahora permite relaciones entre regiones que pertenecen a diferentes tópicos, la posibilidad de introducir salidas de otros clasificadores, etc.

Nuestros experimentos en clasificación y segmentación de imágenes han demostrado el elevado salto en rendimiento que RBLDA supone con respecto, tanto a los modelos tomados como base, como a otras alternativas del estado del arte (incluyendo RBLTM).

Sin embargo, en un escenario no supervisado como el de descubrimiento de conceptos (*topic discovery*), los resultados han sido sorprendentemente diferentes. En este caso, en el que se tratan de detectar conceptos de interés en un conjunto de imágenes no etiquetadas, aproximaciones más simples que RBLDA obtienen mejores resultados. En particular, nuestros experimentos han demostrado que el algoritmo RBLTM constituiría la mejor solución en esta tarea, si bien todo aquél modelo generativo que utilice distribuciones de apariencia más simples mejora los resultados de RBLDA.

Por último, cabe destacar la comparativa ofrecida con respecto a sistemas oficiales evaluados en PASCAL VOC 2010. Es importante destacar cómo estas propuestas se asocian a sistemas complejos que utilizan numerosos descriptores y varios tipos de clasificadores, con lo que resulta bastante complicado realizar comparativas con los algoritmos propuestos (los cuales emplean un número muy limitado de descriptores). Aún así, nuestros resultados son aceptables en clasificación (en torno a la mediana de las propuestas oficiales) y muy buenos en segmentación (por encima del 75% de las propuestas).

Líneas futuras

En esta sección se comentarán las líneas de investigación más prometedoras que surgen como continuación del trabajo realizado en la presente Tesis.

Con respecto al algoritmo de segmentación espacio-temporal, una dirección clara de trabajo es la aplicación del mismo al *tracking* o seguimiento de objetos en vídeo. El mismo modelo probabilístico adaptativo que se emplea para la segmentación iterativa podría adaptarse para el seguimiento de objetos a lo largo de planos consecutivos. Además, esta aplicación requeriría estudiar técnicas de detección de novedad para manejar la entrada o salida de objetos en la escena.

Las ventajas de esta nueva alternativa en el problema de tracking serían variadas: la segmentación se podría refinar mediante la utilización de más de un plano, se podrían estudiar los patrones de movimiento de los objetos a lo largo de una toma y así caracterizar conceptos mediante información espacio-temporal (para clasificación de imágenes, por ejemplo), se podrían utilizar los resultados obtenidos para codificación de vídeo basada en objetos, etc.

Además, como se ha comentado con anterioridad, dado que las medidas de evaluación no son del todo concluyentes, sería interesante aplicar y evaluar este algoritmo a un problema más complejo de recuperación de información multimedia. Si bien el modelo propuesto se ha utilizado en sistemas de análisis vídeo como los enviados a la iniciativa TRECVID 2009 [González-Díaz et al., 2009b], resulta difícil evaluar su influencia en los resultados finales al tratarse de sólo un módulo dentro de un sistema de elevada complejidad.

Los modelos generativos para la representación de imágenes pueden mejorarse de igual modo. La extensión más sencilla consistiría en la utilización de nuevos descriptores sobre el modelo, como descriptores de forma de las regiones. Otra línea más compleja consistiría en la incorporación de modelos de partes [Felzenszwalb et al., 2010b, Felzenszwalb et al., 2010a], pues han demostrado un elevado rendimiento en tareas de detección de objetos.

Atendiendo al modelo teórico de base, nuevos niveles de la jerarquía podrían ser añadidos dando lugar a aproximaciones similares a los *Hierarchical Dirichlet Processes* o HDP [Teh et al., 2006], los cuales explotan correlaciones entre documentos pertenecientes al mismo corpus.

Finalmente, cabe destacar una línea de trabajo muy prometedora: la adaptación de los modelos de tópicos latentes para el modelado espacio-temporal de conceptos en secuencias de vídeo.

ABSTRACT

This PhD. Thesis consists of two well differentiated parts, each of them focusing on one particular field of Computer Vision. The first part of the document considers the problem of automatically generating image segmentations in video sequences in the absence of any kind of semantic knowledge or labeled data. To that end, a blind spatio-temporal segmentation algorithm is proposed that fuses motion, color and spatial information to produce robust segmentations. The approach follows an iterative splitting process in which well known probabilistic techniques such as Gaussian Mixture Models are used as a core technique. At each iteration of the segmentation process, some regions are split into new ones, so that the number of mixture components is automatically set depending on the image content. Furthermore, in order to keep in memory valuable information from previous iterations, prior distributions are applied to the mixture components so that areas of the image that remain unchanged are fixed during the learning process.

Additionally, in order to make decisions about whether or not to split regions at the end of one iteration, we propose the use of novel spatio-temporal mid-level features. These features model properties that are usually found in real-world objects so that the resulting segmentations are closer to the human perception. Examples of spatial mid-level features are regularity or adjacency, whereas the temporal ones relate to well known motion patterns such as translation or rotation.

The proposed algorithm has been assessed in comparison to some state-of-the-art spatio-temporal segmentation algorithms, taking special care of showing the influence of each of the original contributions.

The second part of the thesis studies the application of generative probabilistic models to the image representation problem. We consider "image representation" as a concurrent process that helps to understand the contents in an image and covers several particular tasks in computer vision as image recognition, object detection or image segmentation. Starting from the well-known bag-of-words paradigm we study the application of Latent Topic Models. These models were initially proposed in

the text retrieval field, and consider a document as generated by a mixture of latent topics that are hopefully associated to semantic concepts. Each topic generates in turn visual local descriptors following a specific distribution.

Due to the bag-of-words representation, Latent Topic Models exhibit an important limitation when applied to vision problems: they do not model the distribution of topics along the images. The benefits of this spatial modeling are twofold: first, an improved performance of these models in tasks such as image classification or topic discovery; and second, an enrichment of such models with the capability of generating robust image segmentations. However, modeling the spatial location of visual words under this framework is not longer straightforward since one must ensure that both appearance and spatial models are jointly trained using the same learning algorithm that infers the latent topics.

We have proposed two Latent Topic Models, Region-Based Latent Topic Model and Region-Based Latent Dirichlet Allocation that extend basic approaches to model the spatial distribution of topics along images. For that end, previous blind segmentations provide a geometric layout of an image and are included in the model through cooperative distributions that allow regions to influence each other. In addition, our proposals tackle several other aspects in topic models that enhance the image representation. It is worth to mention one contribution that explores the use of advanced appearance models, since it has shown to notably improve the performance in several tasks. In particular, a distribution based on the Kernel Logistic Regression has been proposed that takes into account the nonlinear relations of visual descriptors that lie in the same image region.

Our proposals have been evaluated in three important tasks towards the total scene understanding: image classification, category-based image segmentation and unsupervised topic discovery. The obtained results support our developments and compare well with several state-of-the-art algorithms and, even more, with more complex submissions to international challenges in the vision field.

AGRADECIMIENTOS

Largo ha sido el periodo de gestación pero, una vez llegado el momento, uno echa la vista atrás y no puede sino recordar los grandes momentos vividos de estos años. Y es de esos momentos de los que nacen mis agradecimientos a todos aquellos que me han rodeado.

Pido perdón por adelantado, a todo aquél que no figure o se sienta incluído en las siguientes líneas. El tiempo para escribir esto es escaso y mi memoria frágil por lo que seguro que me olvido de personas a quienes tengo que agradecer mucho.

Por supuesto, empezaré por agredecer a Fernando, mi tutor de tesis, su esfuerzo y dedicación durante estos años; por haber estado ahí en todo momento, por su disponibilidad y una paciencia encomiable, por saber transmitir la tranquilidad necesaria ante los altibajos anímicos que aparecen en una carrera de fondo como esta.

También quiero agradecer a aquellos otros profesores del departamento que me han ayudado desinteresadamente en ciertos momentos de mi tesis: Jesús Cid, Emilio Parrado, Harold (yo no soy el culpable...); así como aquellos con los que he compartido trabajos de investigación: Jero, Vanessa, Ascen y Carmen.

Uno de los periodos más bonitos y fructíferos de esta etapa ha sido mi estancia en el CDVP de Dublin, bajo la supervisión de Noel O'Connor y Alan Smeaton. A ellos tengo que agradecerles parte de esta tesis doctoral, y a otros como Colum Foley, Tomasz Adamek, James Lanagan, Kevin Mc Guinness y Paul Ferguson, el haberme acogido de forma fantástica y, por supuesto, el haber compartido involvidables tardes y noches de pintas (gracias a tí también: Guinness).

Quiero agradecer también a la gente de Asturias, empezando por los carbayones (hoy en día, algunos ya convertidos en leyendas urbanas más allá del Negrón): Vega, Soidán, Nare, Canal, Patri, Belén, Lorena. Qué grande es siempre volver a la tierra y pegarse un homenaje gastronómico en cualquier lugar recomendado en la "Guía Vega" (Tenemos que ir pensando ya en sacar la primera edición... ¡Tiembla Michelín! Tus días están contados...). También me acuerdo de Tapia, el único lugar en el que alcanzo la tranquilidad plena y desconecto de todos los quehaceres y preocupaciones (será el mar, que tiene esa capacidad de amansar a cualquiera). Parte de culpa la tienen Carre, Delu, Che, Carmen, David, Marta, Toto, Mariana, Surfing, Manu, Vane, Juanín, Jua. Especial mención para tí, Gus, siempre con nosotros, siempre en nuestros recuerdos.

Y a los vallisoletanos, por seguir seforzándose en mantener el contacto a pesar de la distancia; creo que nombrando al Colón ya estáis todos incluidos.

Le debo muchísimo a mis colegas del GPM con quienes he compartido y comparto mi día a día en la Universidad: a Manolo y su cruzada en pos un mundo mejor; al 'abuelo cascarrabias' Sergio; a Azpi y su Excel; a Edu y su cabeza; a Chelus, mi compañero de cigarrillos; a Rubén, Darío, Rosa, Sara, Raúl, ... También a Rocío, quien siempre está ahí para ayudar a todos en lo que sea menester.

Quiero agradecer profundamente el apoyo que he recibido por parte de mi familia. En especial a mis padres, a los que se lo debo todo, por su abnegación, por anteponer siempre mis necesidades a las suyas, por hacerme mucho más fácil todo el camino. También a mi tía, para la que puedo, sin duda, hacer extensivo lo dicho sobre mis padres. A mi hermano y Noemí, mi familia en Madrid, ¡qué bueno es teneros tan cerca! A mi sobri y ahijada Luci, aún no eres consciente de ello, pero no sabes lo importante qué eres: observarte jugar hace desaparecer cualquier problema y es, sin duda, la mejor terapia contra el estrés de nuestras vidas.

También, por supuesto, a Vero. Gracias por haber venido desde tan lejos (en aquél autobús que llegó al pueblo) porque así he tenido la oportunidad, en las Campas, entre cajas de sidra y la orquesta Jerusalén, de conocerte y compartir contigo mi camino.

Iván

I like work; it fascinates me.
I can sit and look at it for hours.

Jerome K. Jerome

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Computer Vision

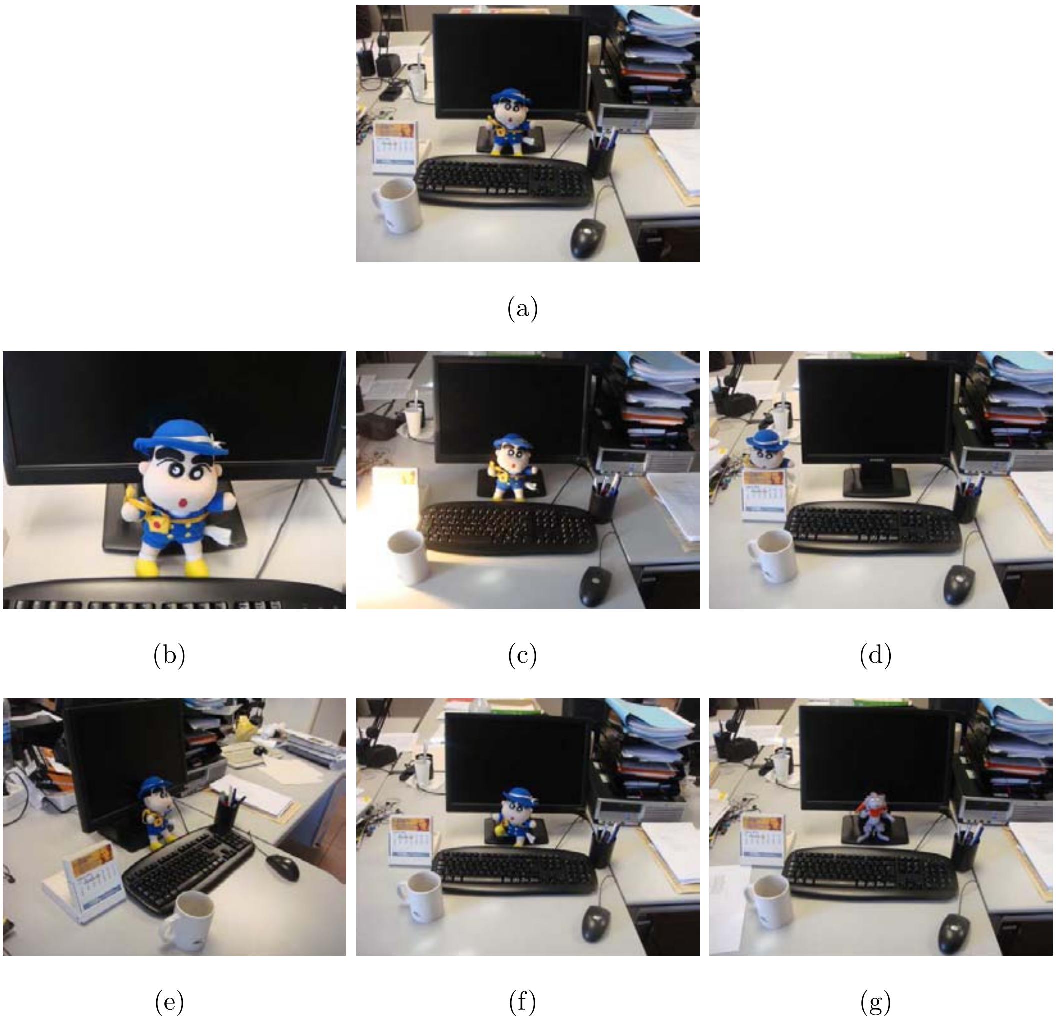
As defined in [Marr, 1982], "Vision is a process that produces from images of the external world a description that is useful to the viewer and not cluttered with irrelevant information". From this definition one question arises: what is a useful description? As discussed by Marr, it clearly depends on the specific purpose of the vision, which is different for each animal in our world. While some animals show a specific purpose in their vision systems, human vision is much more general and complex. However, the presence of several special-purpose mechanisms in our vision, such as suddenly guiding our eyes towards an unexpected movement, endorses this idea. Hence, we can conclude that, since each animal has a specific purpose, the image representation obtained by its visual system should be different.

What we surely know is that, of the five senses, vision is the one that processes most of the data we receive. As estimated in [Davies, 2005], visual information is being captured at our eyes at a rate of about 10 Megabits per second (Mbps). Much of this information is redundant and is compressed by various layers in the visual cortex, so that the higher centers of the brain only interpret a small fraction of the data. However, this amount of information is indeed at least two orders of magnitude greater than the information received by any other sense.

The Computer Vision is the study of methods and techniques whereby artificial vision systems can be constructed and usefully employed in practical applications. It consequently embraces both the science and engineering of vision. [Davies, 2005].

The requirement on the amount of data to be processed becomes a big deal for humans in order to get machines operating as our visual system does. Furthermore, there is an important barrier that still restricts the performance of the machine vision systems. Our brain possesses some 10^{10} cells (neurons), some of which have about 10000 contacts or synapses with other neurons. If each neuron acts as a type of microprocessor, our brain is in fact a supercomputer with a huge number of processors working in parallel and sharing data.

Besides the computational requirements there are other underlying factors that make vision such a difficult task to be carried out by machines. Computer vision systems often face challenges such as scale change, variations on lighting conditions, viewpoint changes, partial occlusion, deformation of non rigid objects or intra-class variation in visual concepts. The reader can find some illustrative examples in Fig. 1.1.



(a)

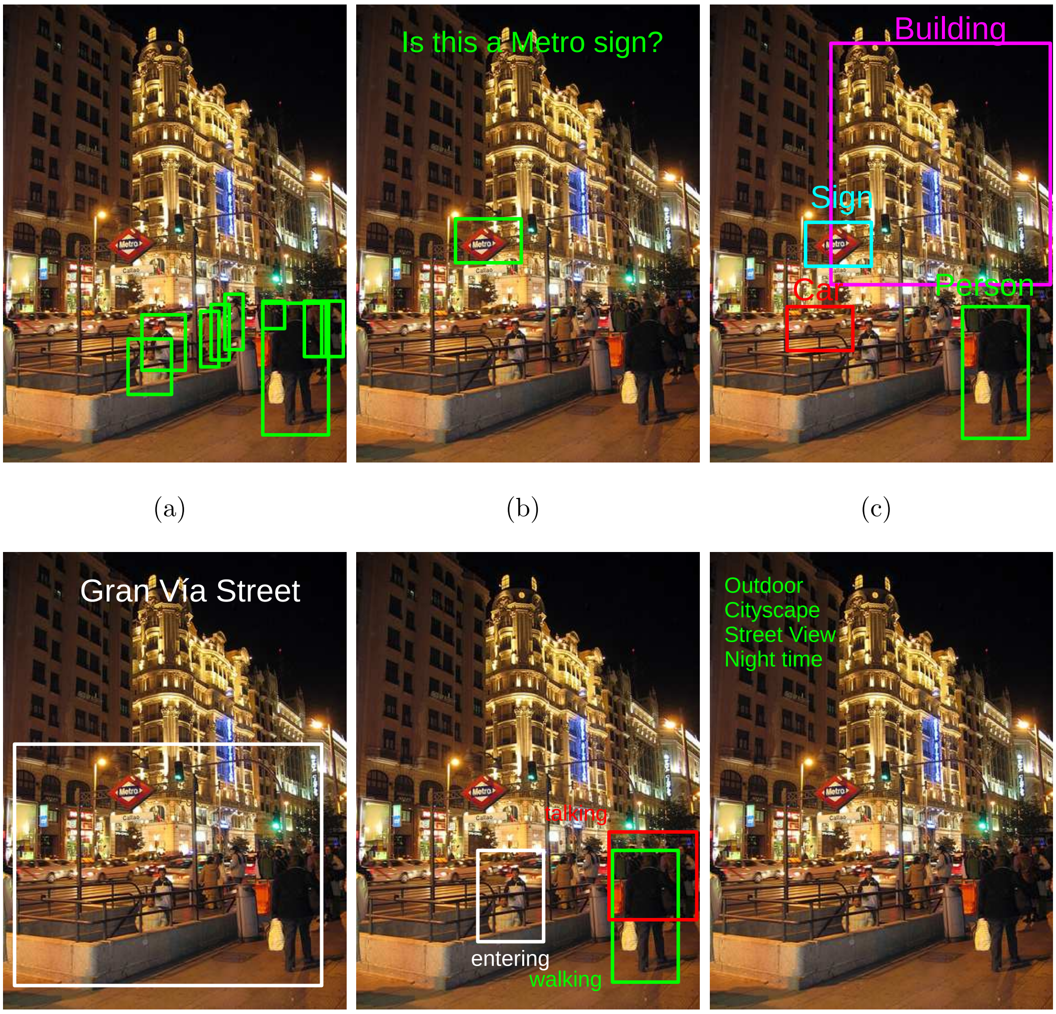
(b) (c) (d)

(e) (f) (g)

Figure 1.1: Some examples of challenges in Computer Vision: a) original image, b) scale variation, c) varying lighting conditions d) partial occlusion e) viewpoint change, f) object deformation and g) intra-class variation.

1.2 Topics and applications in Computer Vision

Computer Vision is an area of research and engineering that covers a wide range of particular tasks. Next, we provide a brief list of topics that have gained much attention by the Computer Vision community:



(a) (b) (c)

(d) (e) (f)

Figure 1.2: Some examples of relevant topics in Computer Vision

- Detection of specific concepts: some expert systems detect instances of specially interesting concepts such as faces, humans or cars. In general, these systems make use of ad-hoc detectors that are specially sensitive to discriminating features of the concept. It constitutes a very traditional topic in video-surveillance, assisted driving, etc. An example in human detection is shown in Fig. 1.2(a).

- Concept verification: in applications where the detection and localization of the intended object is easy, this field is in charge of verifying if a selected region of the image represents the concept of interest. An intuitive example is shown in Fig. 1.2(b).

- Object categorization: Although it might look similar to the previous mentioned detection task, object categorization sets a much more ambitious purpose: following a generic approach (in contrast to ad-hoc), to develop object detectors and classifiers. We add the term classifier since some notion of a multi-class problem is needed in the sense that a region of the image cannot belong to more than one object (besides occlusions or other artifacts). This topic is represented in Fig. 1.2(c).

- Identification: this topic involves retrieving an individual instance of a subject. It differs from object categorization in the sense that identification does not need to generalize the appearance of a concept. It is represented in Fig. 1.2(d), where an individual instance of the concept street, Gran Vía Street, is identified.

- Activity recognition: this field, very common in video surveillance, is in charge of recognizing actions or activities occurring at a scene. Although it might be likely to require motion information (video content) it is often easier to infer actions from static content (images and static features). This topic is illustrated in Fig. 1.2(e)

- Scene description: this topic conceives a scene as a whole and describes its content with a set of general categorizing labels that can be shared across many different images. An illustrative example is provided in Fig. 1.2(f).

- Motion analysis: this field groups several activities related to the motion analysis in video sequences such as the analysis of the camera motion or ego-motion, the tracking of particular objects, or the optical flow computation.
- Image restoration: a particular area in Computer Vision that aims to minimize the effect of different sources of noise from an image: optical blurring, camera shaking, sensor noise, etc.).

Some of these topics are applications by themselves whereas others become parts of more complex systems. As described in [Szeliski, 2011], Computer Vision is being applied to several industrial and consumer-level applications. We next list some of them:

- Optical Character Recognition (OCR): in document scanning or automatic number plate recognition.
- Machine inspection: in industrial environments, machines inspect the state, quality or position of pieces.
- 3D model building: fully automated construction of 3D models from aerial photographs.
- Medical imaging: to perform automatic processing of images of the human body (or parts and function thereof) for clinical purposes.
- Automotive safety: assistance elements that detect known or unexpected elements in the road (traffic signs, pedestrians, etc.).
- Professional content edition: to mix computer-generated elements with real content by estimating a 3D model of the content in the original footage and properly introducing the virtual elements in the scene.
- Motion capture: in computer animation, these systems capture motion from real actors using vision-based techniques.

- Surveillance: probably one of the most active areas that use Computer Vision, not only for human surveillance but also for traffic control.
- Biometrics: for access authentication, Computer Vision and biometrics have jointly evolved for the last decades (fingerprint, earprint).
- Stitching: generating stitched panoramas from overlapping single photos.
- Exposure bracketing: under challenging lighting conditions, to generate a high quality photo by merging photos taken at different exposures.
- Morphing: using morph transitions to convert images into others.
- 3D modeling: to generate 3D models from several 2D snapshots of an object.
- Face detection: in many photo cameras, faces are detected to improve the quality of that specific area of the scene.

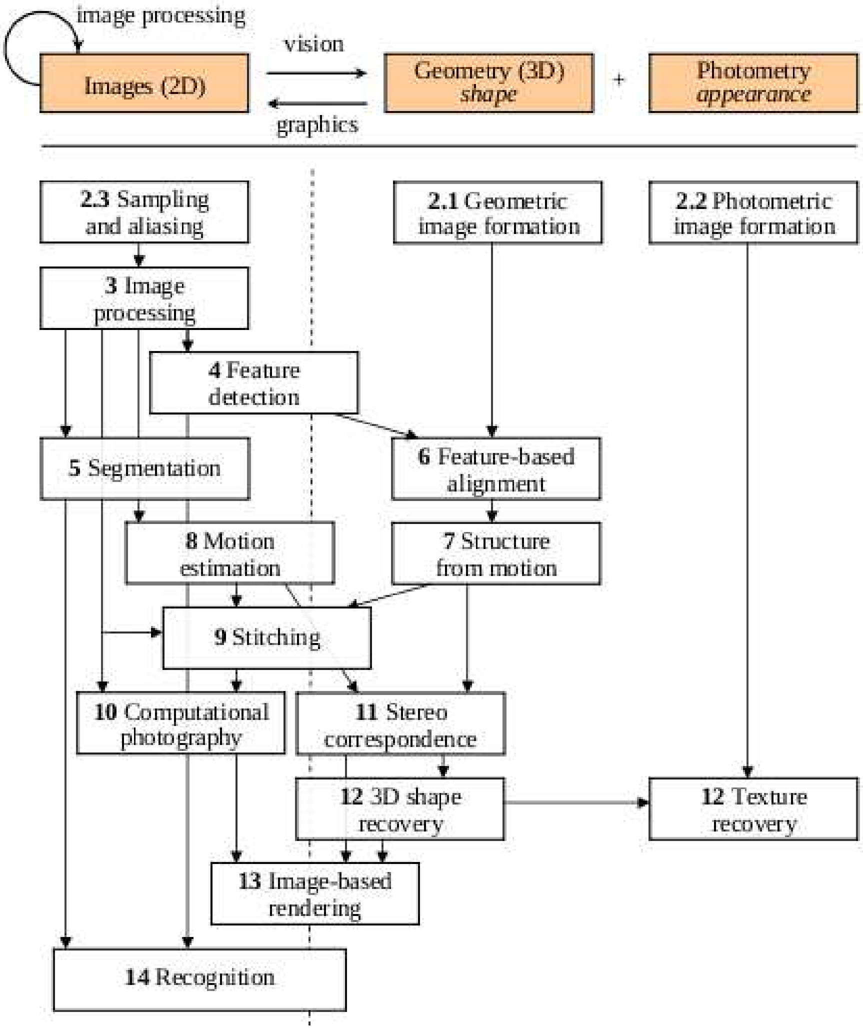


Figure 1.3: Structural representation of the topics involved in Computer Vision. Topics are roughly positioned on the horizontal axis depending on whether they are more closely related to image-based (left), geometric-based (middle) or appearance-based (right) representations. Taken from [Szeliski, 2011]

1.3 Focus of the Thesis

In this section, the specific focus of the Thesis is discussed. In order to establish the main objectives of the Thesis it is useful to locate our contributions and areas of work in the vast field of Computer Vision. Fig. 1.3, taken from [Szeliski, 2011], shows a structural representation of the topics involved in Computer Vision (some of them previously mentioned in section 1.2). Topics are roughly positioned on the horizontal axis depending on whether they are more closely related to image-based (left), geometric-based (middle) or appearance-based (right) representations. This Thesis focuses on two specific topics: Segmentation (5) and Recognition (14). As shown in the figure, both of them locate mainly on the image-based representation and, for the particular case of recognition, on the geometry-based representation to some extent.

Image segmentation is the task of finding groups of pixels that ‘’go together”. In statistics, this problem is known as cluster analysis and is a widely studied area with hundreds of different algorithms [Jain and Dubes, 1988], [Jain et al., 2004]. In Computer Vision, image segmentation is one of the oldest and most widely studied problems. In this Thesis, segmentation is studied from two points of views: a) blind segmentation of images in video sequences, and b) category-based segmentation in images.

Blind image segmentation divides an image into a set of segments that are homogeneous with respect to either one or a combination of similarity measures. It therefore constitutes an unsupervised approach in which no prior knowledge is retrieved about the contents and can be related to the lower levels of the Human Visual System (HVS). In this Thesis, this problem is analyzed on video content, so that motion information is also available to generate the image segmentations.

In contrast, *category-based image segmentation* makes use of higher levels of the HVS in the sense that some prior knowledge about the objects to be segmented is assumed. In this case, the objective is not simply dividing images into a set of homogeneous regions, but associating a group of pixels to one among a set of predefined semantic categories (e.g. aeroplane, sky, car, road, etc.). Hence, this topic is closely related to object recognition and requires performing supervised training with labeled data.

Since recognition still represents a broad topic in Computer Vision, this Thesis particularly focuses on the more specific topic of category recognition. The generic category recognition problem goes beyond the instance recognition problem and proposes the characterization and detection of instances of the same category. Obviously, each instance might show notable differences with others of the same category and might have some individual properties, so that recognition algorithms must be able to both extract features that are consistently associated to the category and learn different instantiations of the same category.

Although the two areas involved in the Thesis might seem to be unrelated, they are, in fact, strongly connected due to the fact that blind segmentation approaches are then incorporated in the proposed models for image representation, category recognition and category-based image segmentation as prior information about the spatial layout of the image. Moreover, category-based segmentation refines these previous blind segmentations and further associates each region to a semantic category.

1.4 Goals, contributions and structure of the Thesis

Two are the main goals of this Thesis: 1) to develop blind image segmentation algorithms for video sequences; and 2) to design probabilistic frameworks for image representation and category-based segmentation. Here we briefly summarize the main content of each chapter:

Chapter 2 aims at developing spatio-temporal segmentation algorithms for video sequences. We first briefly discuss the state-of-the-art on spatio-temporal image segmentation before presenting our contributions. In particular, our proposed solution is based on a well known probabilistic clustering technique: the Mixture of Gaussians. The scenario is as follows: first, a video sequence is divided into shots; then, for each shot, some frames are extracted that help to provide a segmentation of one frame, called the keyframe. In particular, we pursue the following goals:

- The algorithm should fuse two sources of information: spatial (color) and temporal (motion) data obtained from video sequences. The incorporation of temporal information will provide semantically more meaningful segmentations in which spatially heterogeneous regions will be merged if they show to move coherently.
- A robust motion estimation module has to be designed that performs well on varied video content including: varied patterns and magnitude of motion, homogeneous and cluttered scenes, scale variations, etc.
- The probabilistic framework must provide a strategy to split regions into new ones at each iteration. This process, carried out by means of the inclusion of prior distributions, should keep unaltered the remaining regions (those ones that have not been split).
- The segmentation approach will make use of novel spatio-temporal mid-level features that model properties that are found in real-world objects and motion patterns. This kind of features will be used to make decisions on whether add or not new regions during the splitting process to end up with semantically meaningful segmentations.

The obtained segmentation will become the basis for further processing such as object detection, scene recognition or object tracking.

Chapter 3 describes our experiments on the blind segmentation algorithm. A challenging video database has been chosen that has been specifically designed for multimedia information retrieval rather than for image segmentation purposes. Traditional image segmentation databases are normally specially tailored for the segmentation task and do not contain more challenging conditions that are present in real problems. Hence, we sincerely believe that the application of the segmentation algorithms over this database is productive and meaningful, although does not provide so pretty results.

On Chapter 4 we explore probabilistic models for image representation. We have carefully selected the expression "representation" since the proposed models provide useful cues for several Computer Vision tasks such as image classification and object recognition, unsupervised topic discovery and category-based image segmentation. Generative probabilistic models constitute a suitable paradigm for such tasks and, in particular, our proposals are based on well known Latent Topic Models, such as Probabilistic Latent Semantic Analysis [Hofmann, 2001] and Latent Dirichlet Allocation [Blei et al., 2003]. Latent topic models are generative techniques that explain images relying on the Bag-of-Words (BoW) assumption, which considers documents as sets of unordered visual descriptors. Furthermore, they consider some latent variables that are hopefully associated to the semantic concepts that are present in the images.

In this chapter, after reviewing the literature concerning several fundamental aspects, two algorithms are proposed: the Region-Based Latent Topic Model (RBLTM) and the Region-Based Latent Dirichlet Allocation (RBLDA), which:

- Extend basic Latent Topic Models in order to model the spatial distribution of topics (concepts) in images.

- Incorporate previous blind image segmentations and provide cooperative models so that information is shared among regions in an image. This step will provide coherent image representations so that semantically related concepts (e.g. sky/aeroplane) tend to appear spatially connected.

- Enhance the appearance model of basic approach by handling the relations between descriptors that lie in the same region. The original proposals used multinomial or discrete distributions and assumed independence between the image descriptors, what leads to non optimal solutions and poor performance.

- Provide flexible frameworks that are able to manage unsupervised, partially supervised and supervised tasks. Furthermore, two kind of labels are accepted: image-level annotations or tags, and region-based annotations, by means of bounding boxes or pixel-wise ground truth segmentations.
- Develop simple and closed methods for learning and inference. In many cases, some of the expressions are too complex or even untractable; however approximate inference methods and lower bounds will be proposed that optmize model parameters in a feasible way.

On Chapter 5 the proposed models are evaluated in several tasks, either unsupervised (topic discovery) or supervised (classification, segmentation) environments. In particular, official PASCAL VOC 2010 databases [Everingham et al., 2010] will be used so that our results can be fairly compared to other state-of-the-art methods and systems.

Finally, on Chapter 6 we summarize the main contributions of the Thesis, draw our conclusions and outline future lines of research.

Chapter 2

Spatio-temporal image segmentation in video sequences

2.1 Introduction

Automatic video analysis and high-level concept detection systems make use of low-level features to infer annotations of the content which bridge what is known as "the Semantic Gap" [Smeulders et al., 2000]. Considering a keyframe as the information unit, two kinds of features can be extracted: global features and local features. The first ones consider the whole frame to extract a feature vector, while the second are typically based on a segmentation step that divides the frame into spatial regions exhibiting a certain degree of homogeneity. This second case requires robust segmentation systems to produce semantically meaningful regions whose shape, color, texture or motion help to classify or recognize multimedia concepts.

Furthermore, the segmentation of objects in images and video sequences plays an important role in computer vision and multimedia processing, since it lies at the base of many scene analysis approaches. In particular, the segmentation step becomes crucial in applications like content-based image and video retrieval, video tracking [Goldberger and Greenspan, 2006], content-based scalable video coding, perceptual video coding or interactive video.

Classic image segmentation algorithms, such as [Kwok and Constantinides, 1997], usually lead to over-segmentations since the spatial features (color, texture,...) are not sufficient to produce meaningful segmentations in the presence of varying illumination, cluttered backgrounds or complex objects composed of many heterogeneous regions. In these cases, the task of grouping regions to produce objects/concepts is entrusted to higher levels of inference.

In addition, when motion is present, it can be used to perform more meaningful segmentations, which are closer to semantic concepts. Spatio-temporal segmentation algorithms use both types of information (spatial and temporal) to produce coherent organizations of pixels, relying on the assumption that those objects whose motion is different than that of the camera are relevant or important. Examples of this kind of algorithms can be found in [Greenspan et al., 2004] and [Wang et al., 2005].

This chapter describes the proposed blind spatio-temporal segmentation algorithm, an iterative approach that is based on the combination of an adaptive clustering technique and a splitting stage that decides whether or not to add new regions to the partition. Two are the main contributions of this work: 1) the adaptive clustering is performed by a mixture model that successfully handles both prior information from the previous iteration and newly detected regions in the scene and, 2) a decision stage that incorporates two kind of features, low and mid-level features, to decide if new regions should be added to the partition or not. We will demonstrate that mid-level features, by modeling spatial and temporal properties of real world objects, help to produce more meaningful segmentations in which regions are closer to semantic concepts.

The remainder of this chapter is organized as follows: firstly, section 2.2 presents a compilation of the most representative related work in spatio-temporal image segmentation. Then, a detailed description of the proposal and its constituent modules is provided in section 2.3. The assessment of the algorithm is deferred to the next Chapter.

2.2 Related Work

Image segmentation, the process of dividing an image into a set of regions that are homogeneous with respect to certain properties, has been traditionally performed following one of the next two approaches: a) graph based approaches and b) clustering-based approaches.

The former represents an image as a graph in which spatially adjacent pixels are connected. Then, manipulating this graph at several hierarchical levels allows pixels in the lower levels to be grouped into regions at higher levels. Normally, this way of operation leads to what are called as merging algorithms. Of course, the inverse operation is also allowed, but less common in the literature. Good examples of these algorithms can be found in [Kwok and Constantinides, 1997], [Moscheni et al., 1998], and [Adamek and O'Connor., 2007].

Clustering techniques segment images by clustering the feature vectors associated to their pixels. A variety of clustering approaches can be found in the literature: from mixture models, which assume that data is sampled from a set of models and assign each pixel to the most likely model [Greenspan et al., 2006], [Hou et al., 2010] , to spectral clustering methods, based on finding the eigen vectors of affinity matrices [Takacs and Demiris, 2008], [Shi and Malik, 2000] and [Eriksson et al., 2007]. In this thesis we will make more emphasis on the former, since our approach is in fact a mixture model.

With independence on the type of segmentation technique, spatio-temporal algorithms fuse both spatial (static features as color or texture) and temporal (motion) information to obtain perceptually more meaningful segmentations. Hence, a crucial factor on their performance lies in the way they fuse both sources of data. Very initial proposals, such as the one in [Choi et al., 1997], simply perform a weighted linear combination of both elements giving place to a joint similarity measure. However, this combination is very simple and requires to accurately set-up the optimal value of a parameter in the combination. In [Moscheni et al., 1998], two separate hypothesis test are performed for spatial and temporal data that are then combined to make decisions in a region merging algorithm. This approach allows using different hypothesis tests, a likelihood ratio test and a modified Kolmogorov-Smirnov test for spatial and temporal information, respectively.

Other more advanced approaches, such as the ones presented in [Tsaig and Averbuch, 2001] and [Lievin and Luthon, 2004], use generative models such as Markov Random Fields as a fusion method in which incorporating constraints concerning spatial continuity, motion terms and temporal coherence. Other approaches that incorporate heterogeneous sources of data into generative models are [Greenspan et al., 2004] and [Goldberger and Greenspan, 2006], in which a Mixture of Gaussians (MoG) models a set of feature vectors that are a concatenation of spatial and temporal features, or [Lehmann, 2011], in which the authors propose the use of one-dimensional hidden Markov autoregressive models (lines and the columns) in order to reduce the computational burden.

Another interesting point of discussion lies around the features considered in the segmentation algorithm. Traditionally, low-level features concerning both temporal and spatial domains have been utilized to group pixels into regions. Representative features in the spatial domain are color and texture descriptors, as well as, contour information (gradients) that helps to locate the boundaries among regions. In the temporal domain, the most simple features are motion vectors obtained by techniques like optical flow [Ince and Konrad, 2008], [Hu and Li, 2010], or robust block-matching [Gonzalez-Diaz et al., 2007, Gonzalez-Diaz and de Maria, 2007, Gonzalez-Diaz and de Maria, 2008].

An important issue arises due to the fact that these low-level features are often noisy. Situations such as non-uniform illumination cause spatial features to be less reliable. Furthermore, motion vectors are noisy due to the suboptimal results achieved by the state-of-the-art 2D motion estimation methods, that fail in the presence of homogeneous regions or non rigid-motion. More advanced low-level features were proposed to overcome this problem, such as the Displaced Frame Difference

[Choi et al., 1997], that are more robust and stable.

However, the low-level features do not correspond well with high levels of human perception and thus may not lead to perceptually meaningful segmentations. In this context, several authors have proposed the use of mid-level features that model geometric properties of real world objects. If a segmentation algorithm takes advantage of this information, the resulting regions can be closer to semantically meaningful objects. Previous works, such as [Bennstrom and Casas, 2004] and [Adamek and O'Connor, 2007], have successfully introduced the so-called *Syntactic Visual Features*, which model perceptually known geometric properties such as homogeneity, compactness, regularity, inclusion or symmetry.

Extending this idea, well-known motion patterns are usually found in the real world and they have resulted in several parametric motion models: translational, rotation/scaling, affine, perspective or quadratic. These parametrizations can serve to obtain compact motion descriptors either for objects or the camera. These descriptions are also more robust against noise, a serious problem when handling local motion vector maps. In [Aghbari et al., 1998], the authors propose a parametrization for camera and object motion that is used to perform mid-level segmentations and posterior indexing of video sequences.

2.3 Proposed Spatio-Temporal Segmentation Algorithm

2.3.1 Algorithm overview

The algorithm, proposed in proposed [González-Díaz et al., 2008], makes use of the well-known Expectation Maximization (EM) algorithm for Mixtures of Gaussians (MoG). Based on the solution proposed in [Greenspan et al., 2004], this algorithm tries to generate more robust and coherent segmentations using motion information as well as color and spatial features. One of the main advantages of the algorithm is that it is non-parametric in the sense that the number of regions is automatically inferred during its execution.

The whole system involves many other modules, as depicted in Fig. 2.1. For each set of images, which includes a keyframe and a preceding and subsequent frame, the Motion Estimation (ME) module computes the Local Motion Vector Map. Based on this local motion information and the camera motion estimation performed by the Motion Parametrization (MP) module, an initial coarse motion-based segmentation is performed by looking for outliers in the map of motion vectors. This initial segmentation sets the number of classes for the first iteration of the clustering algorithm, K_{init}, but also initializes the MoG. The clustering module uses a splitting technique that performs various iterations with different values of K (the number of clusters), until the K-management module does not find any new regions to be added to the segmentation. The structure and purpose of each of the modules is described in-depth in the next sections.

2.3.2 Motion Estimation Module

The Motion Estimation Module (ME) generates the Motion Vector (MV) Map associated with the keyframe by making use of a temporal window of 3 frames (a preceding, keyframe and next frame). Fig. 2.2 shows some examples of this set of

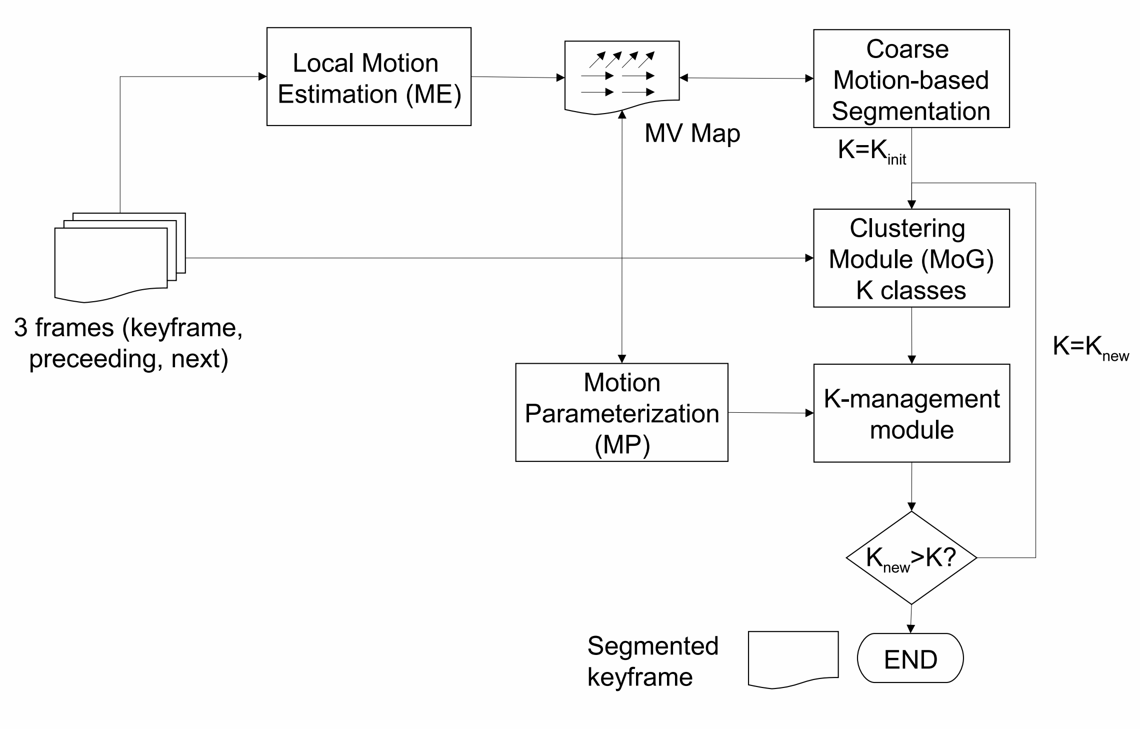

Local Motion Estimation (ME)
MV Map
Coarse Motion-based Segmentation
K=K$_{init}$
Clustering Module (MoG) K classes
3 frames (keyframe, preceeding, next)
K=K$_{new}$
Motion Parameterization (MP)
K-management module
K$_{new}$>K?
Segmented keyframe
END

Figure 2.1: Flowchart of the proposed segmentation algorithm. Three frames feed the system to allow for the motion estimation. Then, only the keyframe is processed by the hierarchical segmentation scheme.

frames. It is worth mentioning that, in order to obtain frames that are different enough to robustly estimate motion, we subsample the video content to 3 frames per second (fps).

Specifically, this module uses a Hierarchical Block-matching Algorithm that computes local motion vectors following a coarse-to-fine approach. A hierarchical approach allows the ME module to obtain better estimations even in the presence of large homogeneous regions. In particular, the algorithm involves N levels (7 in our case) through which the Block Size (BS) decreases logarithmically, starting from a BS of 64 (8x8) pixels (until it reaches a BS of 1). In the proposed implementation, the search range (SR) (dimensions of a rectangle around the initial center point) is set to $SR = (2 * BS) \times (2 * BS)$.

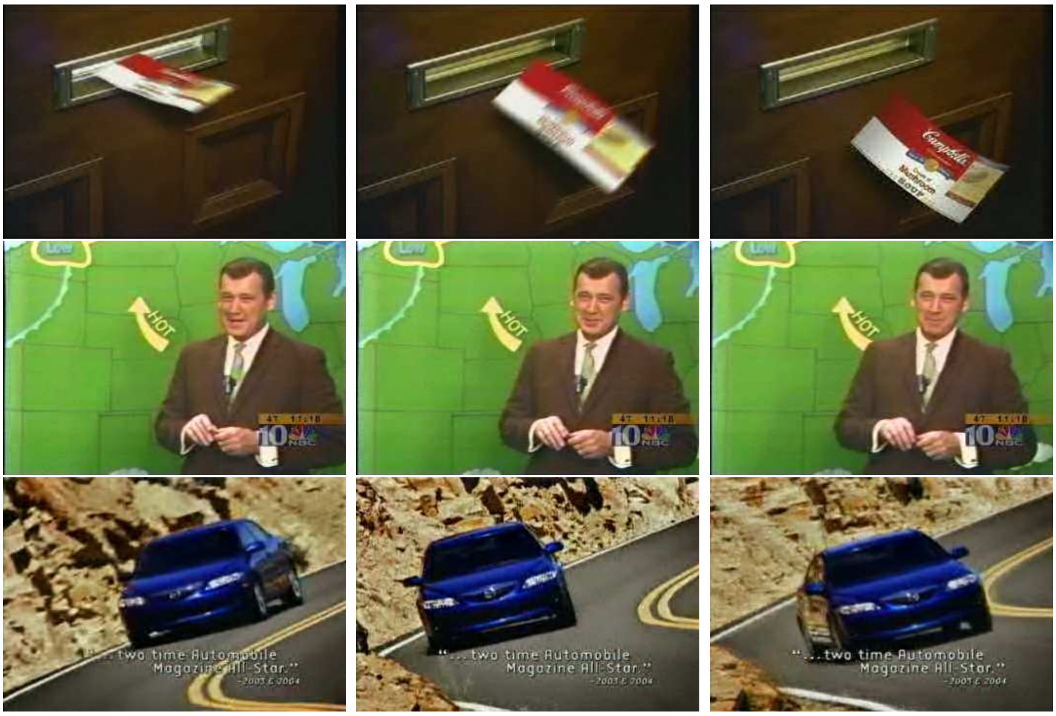


Figure 2.2: Some examples of the frames used to compute motion by the proposed spatio-temporal segmentation algorithm. Previous, current and next frame are shown at left, center and right columns, respectively.

A specific cost function is proposed to strengthen more likely motion patterns. At level l, the cost function obeys:

$$C(\mathbf{mv}^l) = SAD(\mathbf{mv}^l) + \lambda_1|\mathbf{mv}^l - \mathbf{mv}^{l-1}| + \lambda_2|\mathbf{mv}^l| \qquad (2.1)$$

where $\mathbf{mv} = (mv_x, mv_y)$ represents a potential solution for the MV, λ_1 and λ_2 are regularization parameters, and SAD represents the Sum of Absolute Differences between the block being coded I and the block taken as reference I_R. This cost function regularizes the motion vectors by means of the two last terms: the former regularizes the MV with respect to previous estimations at higher levels, while the latter enforces moderated magnitudes for the MVs.

To deal with occlusions, the ME incorporates a temporal window of 3 frames. Since real sequences do not tipically exhibit stationary motion, we have preferred to estimate motion using the preceding frame, and use the next frame only in cases where a potential occlusion is detected. A block i is considered a potential occlusion and, thus, estimated using the next frame, when (1) the cost C_i of the block is higher than an adaptive threshold $C_i > \alpha\bar{C}$, with α being an adjustable parameter and $\bar{C}$ the mean cost of the motion vectors in the frame; and (2) another block in its neighborhood is pointing to it. In this case a ME is performed using the next frame and, if the solution is better, i.e. $C_i^{next} < C_i^{prec}$, the motion vectors are set using the next frame.

Once the MV map has been computed, the MP module allows us to estimate and compensate the camera motion of the sequence, thus producing the final MV map.

2.3.3 Motion Parametrization Module

This module generates Motion Parametrizations based on well-known motion patterns. For simplicity, this module employs a Restricted Affine Transformation (RAT) to model motion as described in [del Blanco et al., 2007]:

$$RAT = \begin{pmatrix} s\cos\theta & s\sin\theta & t_x \\ -s\sin\theta & s\cos\theta & t_y \\ 0 & 0 & 1 \end{pmatrix} \tag{2.2}$$

where s is the scale, θ is the angle of rotation and the vector (t_x, t_y) represents the translation. The motion parameters are estimated through a robust estimation technique based on Random Sample Consensus (RANSAC) [Stewart, 1999] and the Least Median Squares Algorithm. This technique is very robust since it considers the presence of outliers in the set of original points (in our case vectors that have not been properly estimated or regions that move in a different way than the camera). RANSAC estimates the parameters on several small sets of points with the objective that at least one set does not contain outliers. Then, the parametrization that produces the lowest median error in the whole dataset is selected as the final one. The interested reader is referred to [Stewart, 1999] for details.

The objective of this module is twofold. First, it serves to estimate and compensate for the camera motion; and second, it provides region motion parametrizations that are used in the K-management module.

2.3.4 Coarse Motion-based Segmentation

Once the local and camera motion have been estimated, this module looks for outliers in the Restricted Affine Model, thus producing connected regions that serve as an initialization for the first clustering iteration.

Detecting outliers using RANSAC is straightforward since it involves analytical methods to perform this task. Then, a morphological opening serves to generate connected regions that become a practical starting point for the clustering module.

2.3.5 Clustering Module (Adaptive MoG)

As mentioned before, the clustering module is based on a well-known probabilistic framework: the Mixture of Gaussians model. The MoG takes a description of each pixel based on heterogeneous information and provides a coherent spatio-temporal segmentation of the keyframe. For this purpose, the proposed algorithm groups pixels into different clusters (each of them showing some homogeneity) so that a global measure is maximized. The clustering algorithm is totally unsupervised except for the number of classes/clusters K. To avoid this limitation, we propose an iterative approach that allows for automatically obtaining the optimal number of classes in each case.

Feature Extraction

The selected feature space includes information coming from heterogeneous sources such as color, spatial location and motion. Particularly, a 7D feature is defined with the following components: (a) (L, a, b) components of the CIELab color space; (b) spatial coordinates (x, y) of each pixel (to get spatially-coherent segmentations); and

(c) (mv_x, mv_y) components of the motion vectors for every pixel in the keyframe.

After the features are extracted, the components are linearly scaled to produced inputs with zero mean and unitary standard deviation.

Adaptive probabilistic clustering

This section describes the probabilistic model used to group the pixels into different clusters that represent spatiotemporal coherent regions of the keyframe. To this purpose, an extension of the Mixture of Gaussians model is used. In a MoG model the probability density function (pdf) of the pixels x in the keyframe is represented by a mixture of K Gaussians. Then, in our particular approach, the EM algorithm is used to find the values of the parameters that produce the Maximum Likelihood (ML) estimate of the given data x.

Previous developments in the area have proposed the use of priors that represent previous knowledge about the regions to be obtained in the segmentation. In [Goldberger and Greenspan, 2006] the pdf of the pixels obeys:

$$p(x, \theta) = \left[\sum_{k=1}^{K} \alpha_k \mathcal{N}(x|\mu_k, \Sigma_k) \right] p(\theta|\theta_0) \tag{2.3}$$

where α_k are the mixing coefficients of the MoG, $\mathcal{N}$ is a Normal distribution and θ is the parameter set $\theta = \{\alpha_k, \mu_k, \Sigma_k, k = 1...K\}$. Moreover, θ_0 represents the prior knowledge about the parameters. The inclusion of priors makes the EM algorithm to find a Maximum A Posteriori (MAP) rather than a Maximum Likelihood (ML) estimates of the parameters.

In this thesis we propose a new conjugate prior parameter set that uses a diagonal matrix $M_{\beta k}$:

$$M_{\beta k} = diag(\beta_{1k}^{-1/2}, \beta_{2k}^{-1/2}, ..., \beta_{dk}^{-1/2}) \tag{2.4}$$

where $d = 7$ is the dimension of the feature space.

The objective of this matrix $M_{\beta k}$ is to allow for controlling the balance between prior and new models. As mentioned before, an iterative approach is proposed that splits regions into a new set at each iteration. Hence, at a given iteration, the MoG should keep unchanged regions in memory so that their properties remain fixed, whereas parameters of the new regions should get a proper degree of freedom.

The matrix approach manages this balance between models independently for each of the components of the feature space (through β_{ik}). This model extends the previous solution of [Goldberger and Greenspan, 2006], in which a scalar parameter β_k acted as the adaptation coefficient that controlled the balance between known and new/unknown models for each of the classes k (see [Goldberger and Greenspan, 2006] for details). In this thesis, due to the inclusion in the input vector of new features coming from heterogeneous sources, the diagonal matrix $M_{\beta k}$ provides different values β_{ik} for each dimension i of the feature space. The benefit of this approach is twofold:

1. When performing both segmentation and tracking (which is not the case in the experiments presented here), it is important to differentiate among the considered input features. From one frame to another it is more plausible that a region experiences larger changes in its position (in case of translational motion) or shape (rotations, occlusions ...) or motion vectors (non-stationary motion) rather than in its color-related features (only specific concepts like explosions, dramatic occlusions or illumination changes can produce meaningful changes in color). Furthermore, one can decide the level of freedom depending on the type of region and previous frames information.

2. In a hierarchical framework such as the one presented here, in which several iterations of the clustering stage are performed (in order to reach the optimal number of classes), fixing the parameters of those regions that one wants to be unaltered is really useful. On the other hand, those classes that are new in one iteration of the algorithm should receive a higher degree of freedom in order to adjust better to the intended region. Similarly, some aspects like motion can be fixed after the first iteration in those new classes that do not belong to moving regions, while spatial coordinates should receive more freedom to

spread along the intended region.

The following priors are used. The mixing coefficients are jointly Dirichlet, so that:

$$p(\alpha|\alpha_0) \propto \prod_{k=1}^{K} \alpha_k^{(\alpha_{0k}-1)} \tag{2.5}$$

The precision (inverse of the covariance matrix) is Wishart with m_k degrees of freedom:

$$p(\Sigma_k|\Sigma_{0k}, m_k) \propto |\Sigma_k|^{-\frac{m_k-d-1}{2}} \exp\left(-\frac{1}{2}Tr(\Sigma_{0k}\Sigma_k^{-1})\right) \tag{2.6}$$

where Tr stands for the trace operator. Finally, the mean (conditioned by the precision) is normal with a transformation matrix $M_{\beta k}$:

$$\begin{aligned} p(\mu_k|\mu_{0k}, \Sigma_k, M_{\beta k}) \propto |M_{\beta k}\Sigma_k M_{\beta k}^T|^{-1/2} \times \\ \times \exp\left[-\frac{1}{2}(\mu_k - \mu_{0k})^T(M_{\beta k}\Sigma_k M_{\beta k}^T)(\mu_k - \mu_{0k})\right] \end{aligned} \tag{2.7}$$

Since this approach has too many hyperparameters $\theta' = \{\alpha_{0k}, m_k, \Sigma_{0k}, \mu_{0k}, M_{\beta k}\}$, one should set appropriate values to them. Thus, modeling the previous knowledge about the clusters as a MoG with mixing coefficients α'_k and normal distributions $N(\mu'_k, \Sigma'_k)$, hyperparameters can be set to:

$$\begin{aligned} \alpha_{0k} = 1 + \frac{1}{d}Tr(M_{\beta k}^{-1}M_{\beta k}^{-1})\alpha'_k, \quad m_k = \tfrac{1}{d}Tr(M_{\beta k}^{-1}M_{\beta k}^{-1}) + d \\ \mu_{0k} = \mu'_k, \qquad \Sigma_{0k} = (M_{\beta k}\Sigma'_k M_{\beta k}^T) \end{aligned} \tag{2.8}$$

With this modifications, the EM algorithm proceeds as follows. The Expectation step is not affected by the priors so that the probabilities r_{ik} are computed as:

$$r_{ik} = \frac{\alpha_k p(x_i|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \alpha_j p(x_i|\mu_j, \Sigma_j)} \tag{2.9}$$

where r_{ik} is the posterior probability that the input vector x_i was sampled from the kth component of the mixture. The Maximization step applies the new parameter updating equations:

$$\alpha_k = \frac{\sum_{i=1}^{n} r_{ik} + \frac{1}{d} Tr(M_{\beta k}^{-1} M_{\beta k}^{-1})\alpha'_k}{\sum_{j=1}^{K} \frac{1}{d} Tr(M_{\beta j}^{-1} M_{\beta j}^{-1})\alpha'_j + n} \tag{2.10}$$

$$\mu_k = M_A^{-1} \left(\sum_{i=1}^{n} r_{ik} M_{\beta k} x_i + M_{\beta k}^{-1} \mu'_k \right) \tag{2.11}$$

$$\Sigma_k = \frac{\sum_{i=1}^{n} r_{ik}(x_i - \mu_k)^2 + M_{\beta k}^{-1}((\mu_k - \mu'_k)^2 + \Sigma'_k) M_{\beta k}^{-1}}{\sum_{i=1}^{n} r_{ik} + \frac{1}{d} Tr(M_{\beta k}^{-1} M_{\beta k}^{-1})} \tag{2.12}$$

where $M_A = \left(\sum_{i=1}^{n} r_{ik} M_{\beta k} + M_{\beta k}^{-1}\right)$; for simplicity, the operator a^2 represents $a^2 = aa^T$, and n is the number of pixels in the keyframe ($n = H \times W$ dimensions of the image).

Although equations (2.10-2.12) are complex, a fast implementation is straightforward owing to the diagonal nature of M_β, and this approach requires only to set the values of M_{β_k}. In practice, priors for the Gaussians in the first iteration of the algorithm are initialized using means, covariances and mixing coefficients from the coarse motion-based segmentation and, in next iterations, using information from the previous iteration and the K-management module.

2.3.6 K-Management Module: Using mid-level features to determine the number of regions

As mentioned before, we have proposed a hierarchical approach to automatically set the optimal number of regions. After each iteration of the clustering stage, a novelty-detection phase starts. In this stage, the K-management module is a binary classifier that decides if the regions formed by pixels that have been potentially misclassified (i.e. their likelihood is below a threshold) should be added or not. The output of this module consists of a set of new classes (with their prior parameters) to be added in the next iteration of the algorithm. In the case that the output contains no classes the algorithms finishes and the last segmentation becomes the final one. A complete description of this module is provided in the following paragraphs.

Detecting candidate regions

In order to obtain the candidate regions to be classified, the proposed algorithm searches for pixels that show low likelihood values. At this point, likelihood is better than the posterior probability since the prior knowledge about previous iterations of the algorithm is not useful to detect regions that are potentially misclassified. In order to obtain spatially coherent regions from low-likelihood pixels, an over-segmented version of the keyframe is obtained by a fast and simple algorithm such as the Recursive Shortest Spanning Tree (RSST) [Kwok and Constantinides, 1997]. In practice, for every class obtained in the previous iteration of the algorithm (each class is a component in the mixture), the system looks for sub-regions (from the RSST) with a global likelihood value below an adaptive threshold (in our case, the median likelihood of the class). Alternatively, if any class contains two non-connected regions, the bigger one is considered the main region (looking then for internal sub-regions) while the other is automatically set as the candidate.

Once the regions have been labeled as low/high likelihood, the algorithm extracts the connected regions with low values, that become the candidate regions for this class in the next stage.

Problem parametrization

For every candidate region R_{ij} belonging to a class whose main region is denoted as C_i, a region D_{ij} is generated by subtracting R_{ij} from C_i. Concerning the aforementioned three regions, a set of features are extracted to make the decision. The set incorporates two kinds of features: low and mid-level features.

Let us introduce first the Matusita's similarity measure [Matusita, 1955] between two random distributions. Since, in a MoG, the distributions of the classes are Gaussians (and any marginal distribution is also Gaussian), Matusita's measure between some distributions i and j obeys:

$$\chi = \frac{2^{\frac{p}{2}}|\Sigma_i|^{\frac{1}{4}}|\Sigma_j|^{\frac{1}{4}}}{|\Sigma_i+\Sigma_j|^{\frac{1}{2}}}\exp\{-\tfrac{1}{4}(\mu_i-\mu_j)^T(\Sigma_i+\Sigma_j)^{-1}(\mu_i-\mu_j)\} \qquad (2.13)$$

where μ_i and μ_j stand for the means of the distributions i and j, respectively; Σ_i and Σ_j are the covariance matrices; and p is the dimension of the space.

The selected low-level features include:

1. Color-based similarity between R_{ij} and C_i (χ_{color}): The Matusita's measure is employed to compute this similarity between the color marginal distributions of the pixels belonging to R_{ij} ($\rho_{color}(x_{R_{ij}})$) and C_i ($\rho_{color}(x_{C_i})$).

2. Motion-based similarity between R_{ij} and C (χ_{motion}): the same as the previous one, now marginalizing with respect to motion components $\rho_{motion}(x_{R_{ij}})$ and $\rho_{motion}(x_{C_i})$.

3. Absolute candidate region size (S_{abs}): The size in pixels of the region R_{ij} is normalized between $[0, 1]$ with respect to the dimensions of the image.

4. Relative candidate region size (S_{rel}): The size in pixels of the region R_{ij} is normalized between $[0, 1]$ with respect to the size of the class C_i.

5. Internal/External region (B_{IE}): as commented above, if C_i contains more than one non-connected regions, a binary input is set to 0 if R_{ij} is internal to the main region and to 1 if it is external.

In parallel, some mid-level features are also extracted to provide perceptually meaningful information about the regions, namely:

1. Adjacency (Adj): as stated in [Adamek and O'Connor, 2007], real world objects tend to be compact, thus exhibiting adjacency of their constituent parts. Given R_{ij} and D_{ij}, this input provides a useful information about geometrical relations between the regions:

$$Adj = 1 - \frac{l_{R_{ij}D_{ij}}}{min(l_{R_{ij}}, l_{D_{ij}})} \tag{2.14}$$

   where $l_{R_{ij}D_{ij}}$ is the length of the common boundary between R_{ij} and D_{ij}, and $l_{R_{ij}}$ and $l_{D_{ij}}$ are their perimeter lengths. Values close to 0 imply that the regions are strongly adjacent and viceversa.

2. Regularity (Reg): this compares the complexity of the boundaries of the regions C_i, R_{ij} and D_{ij}. Given the area a_i of a region, its complexity x_i can be measured as the ratio between its perimeter length l_i and the square root of its area a_i: $x_i = l_i/\sqrt{a_i}$. Then, the Regularity is computed as follows:

$$Reg = \frac{x_{C_i}}{\left[\frac{a_{R_{ij}}x_{R_{ij}}+a_{D_{ij}}x_{D_{ij}}}{a_{R_{ij}}+a_{D_{ij}}}\right]} \tag{2.15}$$

   In this case, low values of Reg imply that the complexity of the C_i is quite lower than the resultant complexity after the splitting process, thus recommending not to add a new region.

3. Motion Parametrization Error (MPE): this measure utilizes a parametrization of the motion of C_i computed in the MP module. The MP module employs the Restricted Affine Transformation (RAT) described in eq. (2.2) to generate a motion parametrization for the region C_i. Then MPE is the global Mean Square Error (MSE) between the estimated local vectors (from the ME module) with respect to the values of the parametrization. For each position (x, y) in the region C_i, the square difference between the local vector $mv(x, y)$ and the parametrization $mv^P(x, y)$ is computed in order to calculate the MSE. Large values of MPE are associated to hardly recognizable motion patterns or noise.

4. Motion Model Adjustment (MMA): the MMA computes the MSE of the points of R_{ij} with respect the parametric model obtained for the whole region C_i. If the level of adaptation is low, the region is moving and following a different motion pattern, so that it should be added.

Classifier design

A Multilayer Perceptron (MLP) [Rosenblatt, 1962] with one hidden layer was used to classify each case $\{ADD, DON'TADD\}$. The number of neurons, 5 in our case, was experimentally selected by means of a cross validation process.

Chapter 3

Experimental results on spatio-temporal image segmentation in video sequences

3.1 Experimental Setup: database and performance measures

This sections assesses the proposed spatio-temporal segmentation algorithm that has been previously introduced in Chapter 2.

In order to evaluate the performance of the algorithm a ground-truth (GT) segmentation database has been created. This database contains 120 heterogeneous triplets of images from the news video contents of Trecvid 2006 [National Institute of Standards and Technology, 2006]. The database is available at http://www.tsc.uc3m.es/∼igonzalez/, and contains real images which are difficult to segment using simple features such as color. Each sample from the database contains three images (keyframe, preceding and next frame) and a ground-truth segmentation of the keyframe. Although the ground-truth segmentation has been created by humans and is therefore subjective, some principles have guided its generation: connected regions that are moving coherently are merged independently of their color (this allows for segmenting objects as simple regions when they show different colors, illuminations,...). On the other hand, regions that do not show motion, are segmented by color. Figure 3.1 shows five examples of the Ground Truth database.

The use of the database is as follows: 50 triplets have been used to train the decisor of the K-Management module, 20 to perform a cross-validation with various initializations of the MLP described in subsection 2.3.6, and the last 50 form the test set. To manually segment each of the classes, a scribble-driven semi-automatic segmentation tool has been used (see [McGuinness et al., 2006] for a reference).

Furthermore, with the objective of providing numeric results an integrated image segmentation framework has been used [McGuinness et al., 2007], which implements three different methodologies for evaluation, namely:

1. Berkeley Evaluator [Martin et al., 2001]: this evaluation method includes two measures: Global Consistency Error (GCE) and Local Consistency Error (LCE). Although these measures correspond well with human perception they are not sensitive to over and under-segmentation.

2. Huang Dom Evaluator [Huang and Dom, 1995]: this evaluator computes the Hamming Distance (HD) between intersecting regions. However, in order to be less sensitive to under and oversegmentation, this measure removes the largest regions in GT and query segmentations from the computation. For convenience 1-HD is used in our experiments so that, for every measure, values close to zero correspond to better segmentations.

3. Simple Evaluators based on counting pairs using the Rand Index (RI), the Jaccard Index (JI) and the Fowlkes and Mallows Index (FMI).

This framework has been used to perform two different evaluations: a) a *region based evaluation*, which compares segmentations using the aforementioned indexes

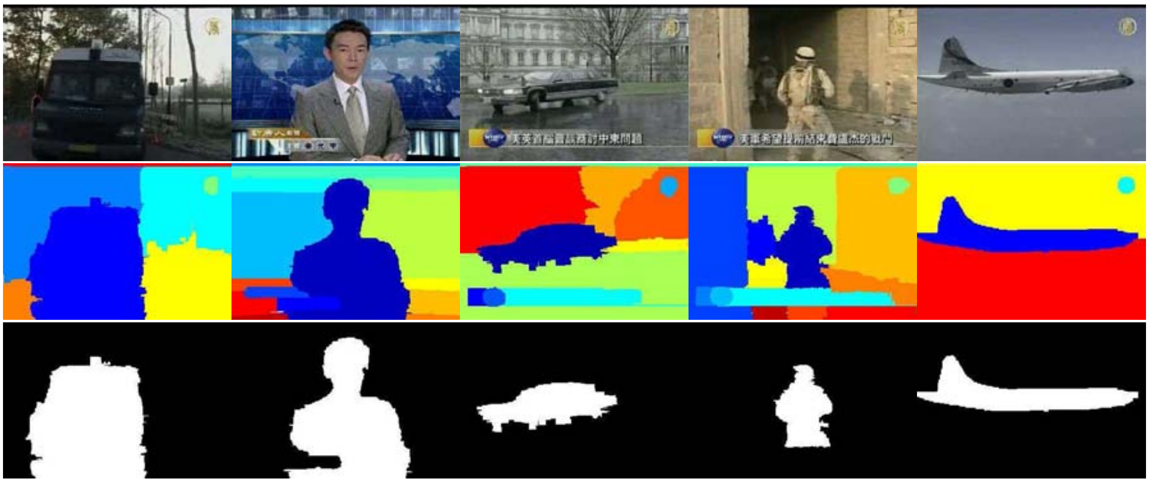


Figure 3.1: Some examples of the developed database including keyframes (top), ground-truth segmentations (middle) and regions of interest (bottom).

and assigns the same influence to every region, and b) *a region-of-interest based evaluation*, which distinguishes among those regions which are considered to be of-interest and others. As proposed in [Ge et al., 2006], a mask is generated for each region of interest and the evaluation measures are computed against a query segmentation in which those regions that show at least the 50% of their points belonging to the region of interest remain unaltered, whereas the rest of the regions are merged to form the background. However, in order to penalize the oversegmentation, each foreground class is considered separately, thus producing a multi-class segmentation, and not a binary mask as suggested in [Ge et al., 2006]. Figure 3.1 shows some examples of ground-truth segmentation and regions of interest in the developed database.

3.2 Experimental Results

In order to assess the performance of the proposed spatio-temporal segmentation algorithm, we have established a meaningful comparison against several reference algorithms, namely:

- MDL spatio-temporal segmentation algorithm (MDL): this approach uses the Minimum Description Length (MDL) criteria to select the optimal number of classes in a MoG approach for segmentation (see [Greenspan et al., 2004] for details). Except for the MDL criteria, the rest of the algorithm is exactly the same as the proposed one.

- Low-level spatio-temporal segmentation algorithm (LLF): an implementation of our proposal in which the mid-level features have been removed when making decisions in the K-management module.

- A Recursive Shortest Spanning Tree (RSST) with spatial mid-level features [Adamek and O'Connor., 2007]: this region-merging algorithm proposes a binary partition tree that uses static features of the keyframe (it does not consider motion information) and incorporates spatial-mid level features to make decisions in the merging process. This approach has been included in the experiments in order to measure the influence of the temporal features at both granularities (low and mid-level).

Table 3.1 shows average results (with standard deviations) for all the considered algorithms in our tests. In addition, Table 3.2 includes results oriented to a region-of-interest-based evaluation. Several interesting conclusions can be drawn from these results:

- The MDL criteria performs poorly when compared to a classifier-based solution. The rationale behind is that classifier-based approaches are trained using a labeled set with ground truth segmentations. Since these segmentations have

Table 3.1: Comparative performance evaluation among the involved algorithms. Values closer to zero correspond to better segmentations. Results on the test set are shown using mean and standard deviation ($\mu \pm \sigma$).

Alg.	**GCE**	**LCE**	**1-HD**	**RI**	**JI**	**FMI**
MDL	0.28 ± 0.10	0.22 ± 0.07	0.38 ± 0.05	0.23 ± 0.09	0.78 ± 0.08	0.60 ± 0.09
LLF	0.26 ± 0.09	0.20 ± 0.07	0.29 ± 0.08	0.28 ± 0.11	0.62 ± 0.13	0.44 ± 0.11
RSST	0.29 ± 0.12	0.19 ± 0.09	0.28 ± 0.08	0.28 ± 0.11	0.62 ± 0.15	0.45 ± 0.14
Proposed	0.25 ± 0.10	0.19 ± 0.07	0.28 ± 0.08	0.25 ± 0.10	0.62 ± 0.14	0.43 ± 0.12

Table 3.2: Comparative of region-of-interest based performance. Values closer to zero correspond to better segmentations. Results on the test set are shown using mean and standard deviation ($\mu \pm \sigma$).

Alg.	**GCE**	**LCE**	**1-HD**	**RI**	**JI**	**FMI**
MDL	0.14 ± 0.09	0.06 ± 0.04	0.13 ± 0.07	0.31 ± 0.11	0.35 ± 0.15	0.21 ± 0.11
LLF	0.04 ± 0.08	0.03 ± 0.05	0.12 ± 0.07	0.32 ± 0.14	0.34 ± 0.15	0.20 ± 0.09
RSST	0.10 ± 0.08	0.05 ± 0.04	0.12 ± 0.07	0.33 ± 0.12	0.36 ± 0.15	0.22 ± 0.10
Proposed	0.05 ± 0.07	0.03 ± 0.04	0.12 ± 0.07	0.29 ± 0.14	0.32 ± 0.15	0.19 ± 0.09

been generated following the same process as those ones in the test set, it is expected that the classifiers approach better to the ground truth segmentation than a general MDL criteria. However, this issue also reflects that a statistical measure such as the MDL does not correspond well with human visual perception. In practice, it is easy to notice that MDL usually generates over-segmented versions of the images (see fourth row in Fig. 3.2 and Fig. 3.3).

- RSST only considers spatial features of the keyframe so that it cannot model objects that, although moving coherently, are composed by spatially heterogeneous regions. However, as seen in Table 3.1, it surprisingly provides results that are more accurate than the those ones achieved by MDL (a spatio-temporal segmentation algorithm) and very close to those ones by LLF and our proposal. This performance decreases if we study the region of interest oriented evaluation (Table 3.2). An interesting rationale for this issue can be found looking at the visual results (sixth rwo on Fig. 3.2 and Fig. 3.3). In these images, background elements (no motion or camera motion) results are very close to our proposal whereas foreground objects (see region-of-interest segmentations) are normally decomposed into several homogeneous regions. Since the proportion of non-interest regions in an image is notably high, this drawback remains hidden in a general evaluation and arises when focusing on the region-of-interest.

- The LLF and our proposal get more similar results; however, the use of mid level features provides a slightly better performance as shown in both evaluations. In general, the proposed solution tends to discriminate better whether to add or not new regions to the segmentation.

With respect to the evaluation measures, it also worth discussing some observations:

- The standard deviations are very high in all cases so that the performance is not very stable along the whole database. This is not surprising since the database shows very challenging shots with a great variety of motion content (slow/fast, different patterns as zoom or pan, and more than one object moving in a scene).

- In addition, none of the measures provides very distinctive performances. Looking at the averages it is difficult to get significantly different results among the different algorithms even when segmentations seem to be very different. We can then conclude that the measures themselves do not model properly human perception.

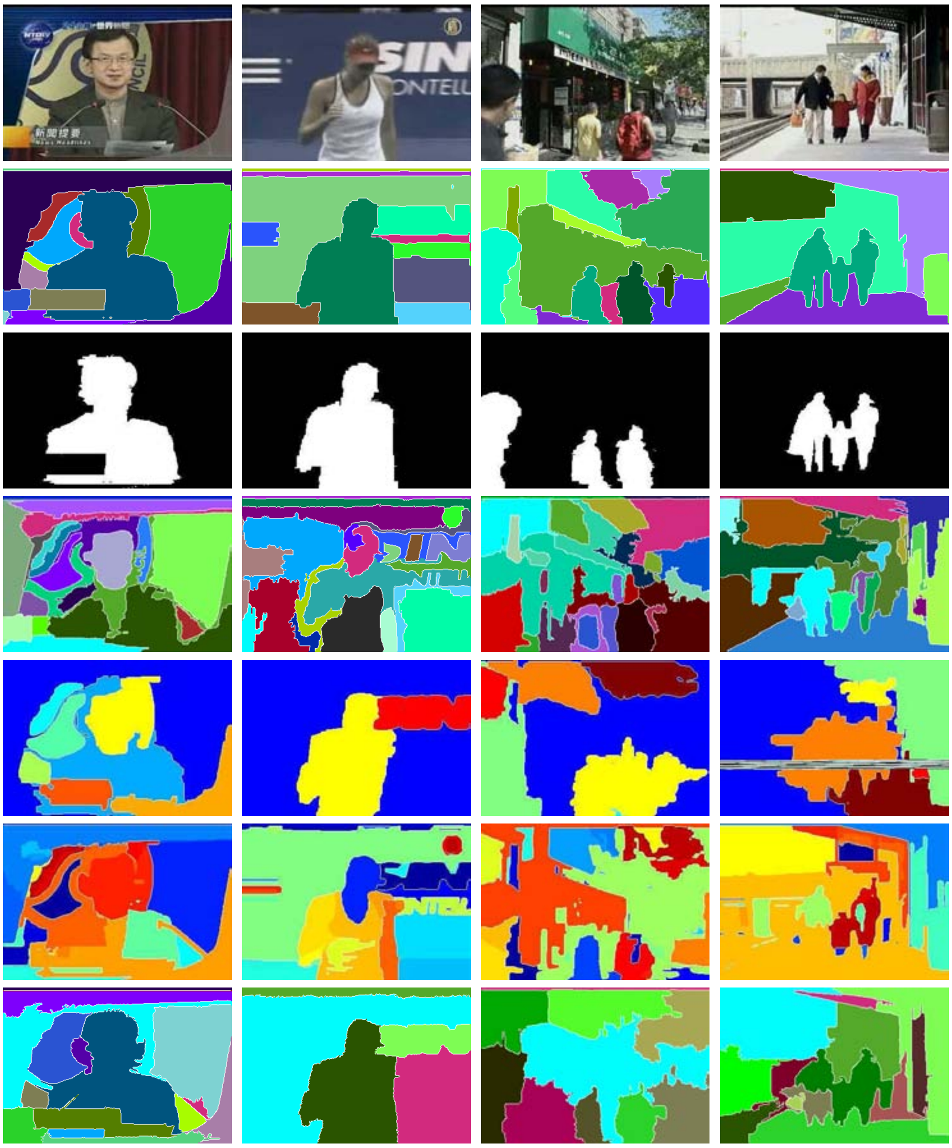


Figure 3.2: Some examples of segmentation results including, from top to bottom, keyframes (first), ground-truth segmentations (second), region-of-interest ground truth segmentations (third), MDL results (fourth), LLF results (fifth), RSST results (sixth) and proposed algorithm results (last)

.

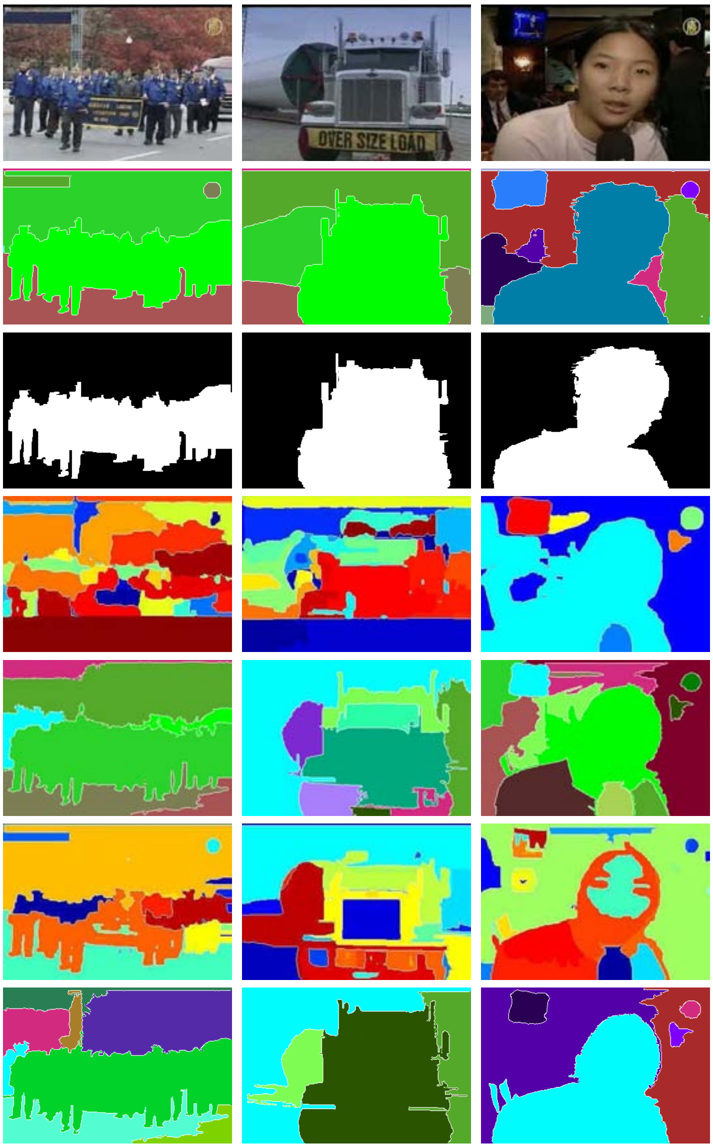


Figure 3.3: Some examples of segmentation results including, from top to bottom, keyframes (first), ground-truth segmentations (second), region-of-interest ground truth segmentations (third), MDL results (fourth), LLF results (fifth), RSST results (sixth) and proposed algorithm results (last)

Furthermore, the use of a hierarchical scheme to get the optimal number of classes provides a considerably reduction on the number of iterations of the algorithm. Our tests have resulted in 3.12 mean iterations for the proposed algorithm in comparison to the 15.46 needed by the clustering algorithm with the MDL criteria.

Finally, it is also noteworthy that the proposed segmentation approach has been successfully applied in a multimedia information retrieval system in the TRECVID project [González-Díaz et al., 2009b]. The results achieved by our system ranked between the median and the 25% percentile that year. However, the segmentation module was embedded in a complex system in which the contribution of this subsystem was not particularly assessed.

3.3 Conclusions

In this chapter we have assessed the performance of the proposed spatio-temporal segmentation algorithm by establishing a meaningful comparison with state-of-the-art segmentation approaches and some variations of the proposed one that help to evaluate the influence of several internal modules. Although the results are not very conclusive in the sense that performance measures do not seem to be discriminative enough, our proposal consistently outperforms all other considered algorithms. Two are the main reasons of this improvement:

1. The iterative adaptive MoG approach that splits regions at each iteration and manages the balance between prior models and new ones.

2. The use of mid-level features that successfully help to make decisions about the splitting process.

Chapter 4

Generative models for image representation

4.1 Introduction

In recent years, a lot of research has been devoted to the image classification, object class image segmentation, and topic discovery problems since they have become necessary parts in contemporary scene understanding systems, which have emerged as a natural extension of the classical image classification and recognition systems. This new approach considers images as collections of semantic objects and, therefore, it should be able to detect and label local regions using a set of semantic classes.

The remainder of this chapter is as follows: in order to properly place our proposal in such a vast field, the next subsections describe the state-of-the-art concerning image classification, object class image segmentation and topic discovery. Then, traditional approaches that serve as the basis for the proposed methods are discussed such as the bag-of-words model and the latent topics models. Once the proper background has been described, two novel generative models are introduced, the Region-Based Latent Topic Model (RBLTM) and the Region-Based Latent Dirichlet Allocation (RB-LDA), that aim to extend basic Latent Topic Models for the considered tasks.

4.1.1 Image classification

Although, initially, the image classification task required systems to classify images among a predefined set of categories, nowadays, these systems are required to detect the presence of different objects/concepts in an image. This new vision of the problem allows images to contain more than one object and thus to belong to more than one category. It is also noteworthy that this task does not require to accurately locate the concept in the image, but simply detect its presence.

For this particular problem, Bag-of-Words (BoW) models have shown exceptional performance and constitute the most prevalent approach. These models were initially proposed for text retrieval and later used in Computer Vision, where the traditional "document" became an image and the "words" were associated with visual words that describe the content of local patches. BoW models make a simplifying assumption on the data distribution in a image, which is simply considered as an unordered collection of visual words. Good examples of BoW models can be found in both discriminative [Wallraven et al., 2003] [Grauman and Darrell, 2005] and generative frameworks [Hofmann, 2001] [Blei et al., 2003]. Originally, these models did not take into account the spatial location of the visual words, what, obviously, limited their performance. More recently, some spatial constraints have been proposed for BoW models to benefit from spatial discrimination to some extent. In particular, the discriminative approach called Spatial Pyramid Matching ([Lazebnik et al., 2006],[Bosch et al., 2007],[Varma and Ray, 2007]) attains improved classification performance by computing image histograms at different spatial levels and a weighted kernel that sets the relative importance of each spatial scale.

4.1.2 Object class Image-segmentation

The *object class image-segmentation* task differs from the automatic blind image segmentation problem that has been previously described in this thesis. Specifically, object-class image segmentation is a technique that, not only divides the image into a set of coherent regions, but also labels these regions according to their category. This new approach requires systems to be trained with labeled data in order to provide semantically meaningful segmentations.

The approaches found in the literature for this problem are diverse. Starting with generative approaches, in [Larlus et al., 2010], BoW, Dirichlet Processes (DP), and Random Fields are combined to provide a non-parametric DP mixture with spatial regularization for object class image segmentation. However, the appearance model is decoupled from the rest of modules and trained separately. In [Lee and Grauman, 2010], a set of unlabeled images is segmented by generating a set of region clusters, representing each cluster by its ensemble (thus modeling intra-class variation), and applying a graph cut that operates on the distances among regions and cluster ensembles.

Among the discriminative approaches, cooperative probabilistic models such as Conditional Random Fields (CRF) are one of the most prevalent techniques. In [Galleguillos et al., 2008], the authors propose a BoW-based appearance model and a CRF-based spatial/context model for object categorization and image segmentation. However, as it happened in [Larlus et al., 2010], the two models (appearance and spatial/context) are trained separately. In [Shotton et al., 2009], CRFs are used for integrating color, texture, location and context into a unified framework. In [Gould et al., 2009a], Gould et al. propose hierarchical models involving appearance and spatial context. This model is further refined in [Gould et al., 2009b] by considering object-detection features.

4.1.3 Topic discovery

The *topic discovery* task requires algorithms that unsupervisely detect topics of interest (concepts) in a set of unlabeled images. Latent topic models, such as the well known Probabilistic Latent Semantic Analysis (PLSA) [Hofmann, 2001] and Latent Dirichlet Allocation (LDA) [Blei et al., 2003] are excellent examples of successful unsupervised algorithms for this task. Both PLSA and LDA are generative models that consider documents as mixtures of latent topics that govern the occurrence of words. The main difference between the two formulations lies in the fact that LDA additionally learns the prior distributions of the topics.

The interested reader is referred to [Tuytelaars et al., 2010] for an interesting survey that compares these and other methods applied to the specific problem of topic discovery.

4.2 Related work on image representation

In this section we provide a review of the related work concerning methods for image representation. In order to properly introduce the proposed Latent Topic Models, we will start with the basic bag-of-words representation, which has become the main approach in the image classification task, mainly due to its simplicity and good performance. Then, Latent Topic Models that make use of this representation will be presented and two well-known approaches will be discussed in depth. Finally, we will review some extensions of the original topic models that consider the spatial structure of visual documents.

4.2.1 The Bag-of-Words Representation

The bag-of-words (BoW) is a very simple approach that treats images as collections of regions that are described only by their appearance, thus ignoring their spatial location and, in consequence, the global image structure. Those models were initially proposed by the text analysis community and later applied in the Computer Vision field by [Sivic and Zisserman, 2003], [Csurka et al., 2004], [Fei-Fei and Perona, 2005] and [Sivic et al., 2005]; or [Snoek and Worring, 2009], in which the application of the BoW to the content-based video retrieval is studied.

The objective of this section is twofold: first,to present the main steps involved by the BoW representation; and second, to describe, for each step in the BoW, several interesting techniques found in the literature. The rationale behind is that BoW representation is a general methodology that does not define how to implement each of the steps in the model, thus leaving many decisions open such as the feature detectors and descriptors, the generation of the word-histograms, the kind of classifiers, etc.

Although several slight variations may exist among different implementations of the model, we briefly describe the main steps of the bag-of-words approach:

1. Detection and description of image patches

2. Visual vocabulary construction

3. Generation of the BoW representation

4. Classification based on the BoW representation

Detection and description of image patches

During the last decade, local patches and their associated descriptors have received a lot of attention from the visual computer community due to their robustness against occlusions and cluttered background. The idea behind of the use of these descriptors is to detect salient points in an image and describe the local regions around them. This process involves two differentiated steps: a) detecting the salient regions; and b) computing local descriptors.

In order to detect salient regions in an image, there are several interesting measures in the literature that have given place to specific detectors. In general, one of the desirable properties for local regions and their descriptors is that they should be repeatable. This means that if we apply a transformation over an image, their corresponding local descriptors should be found in the transformed image and the descriptors should be exactly the same. This idea has lead to the design of descriptors with invariance against certain affine transformations (rotation, scale) or illumination changes. However, it is noteworthy that, in general, the more repeatable is the region/descriptor pair, the less discriminative it is. Well known detectors found in the literature are Harris-Affine [Harris and Stephens, 1988][Mikolajczyk and Schmid, 2002], Hessian-Affine [Mikolajczyk and Schmid, 2002], Maximum Stable Extremal Regions (MSER) [Matas et al., 2002], Difference of Gaussians (DoG) [Lowe, 2004], Laplacian of Gaussians (LoG) [Haralick and Shapiro, 1992], etc. A good discussion about several of these detectors can be found in [Mikolajczyk et al., 2005] and [Tuytelaars and Mikolajczyk, 2008].

In Fig. 4.1 we show some examples of local regions detected by three detectors:

Harris-Affine, Hessian-Affine and MSER.

In what concerns to local descriptors, they should represent local neighboring information around the detected point. They usually describe texture or color in the area defined by the elliptical region (as those shown in Fig. 4.1). Among the proposed descriptors in the literature, the Scale Invariant Feature Transform (SIFT) [Lowe, 2004] deserves a special mention due to the fact that it has shown great discriminative power and has strengthened the use of this kind of approaches.

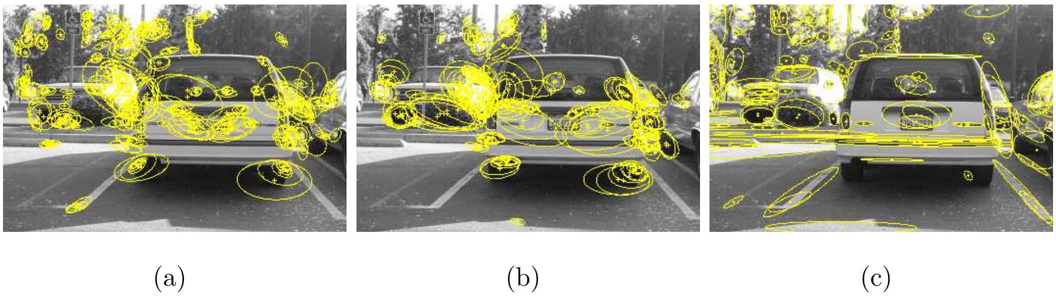


Figure 4.1: Some examples of local detectors. a) Harris-Affine; b) Hessian-Affine; c) MSER. Each ellipse represents a detected region with its orientation and scale in both axis.

SIFT is an algorithm to detect and describe local regions in images. The proposed detector in SIFT is the Difference of Gaussians (DoG) detector, that searches for scale-space extrema. The SIFT descriptor is invariant to image translation, scaling, and rotation, partially invariant to illumination changes, and robust to local geometric distortion. The descriptor is a histogram of spatial gradients on a local region. The gradient at each pixel is considered a tridimensional feature vector, composed by the pixel location (relative to the region center) and the orientation of the gradient. In order to gain some spatial discrimination and model the local structure, the intended region is first divided into a grid of 4x4 cells that give place to 16 histograms of oriented gradients (with eight orientations). Then, the 16 histograms with 8 orientations are concatenated to generate a feature vector of dimension 128. Furthermore, the feature vector is normalized to unit length, and a spatial Gaussian weighting function is also considered to give more emphasis to pixels that are closer to the center of the region. Fig. 4.2 shows the original illustration of the SIFT descriptor by [Lowe, 2004]. For simplicity, in this figure only 2x2 cells are shown in which histograms of oriented gradients are computed.

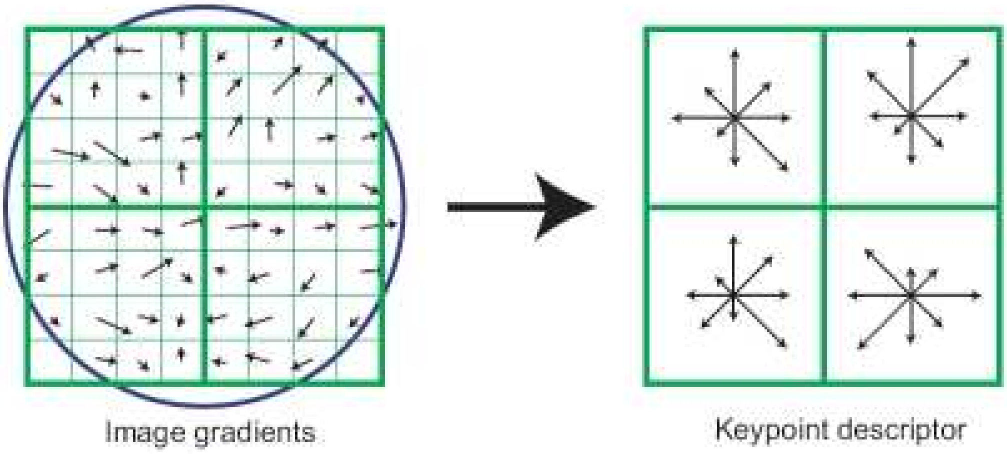


Figure 4.2: Example for SIFT feature construction. Each cell gives place to an 8-bin histogram of gradients orientations. Figure taken from [Lowe, 2004]

Other descriptors of interest that can be found in the literature are SURF (Speeded-Up Robust Features) [Bay et al., 2008], DART [Marimon et al., 2010] or HOG (Histogram of Oriented Gradients) [Dalal and Triggs, 2005].

Visual Vocabulary construction

Since, in general, both the number of local interest regions and the dimension of the local descriptors are high, working on this feature space becomes unfeasible for any machine learning approach. To overcome this issue, the idea of constructing visual vocabularies arises with the objective of assigning each vector to the most similar one among a predefined set of potential values (vocabulary of visual words). This approach can be seen as a vector quantization process that assigns each vector to one of the vectors in the visual vocabulary.

To that end, visual vocabularies are intended to contain representative feature vectors that are repeatable along all documents in a corpus. Hence, clustering techniques are suitable approaches to divide a large dataset into several representative clusters. Most clustering algorithms are based either on iterative square error partitioning or on hierarchical schemes. Square-error-based partitioning algorithms attempt to obtain a partition of the feature space that minimizes some error measure, whereas hierarchical approaches represent the dataset in a hierarchical structure, like a tree, following some heuristics. Examples of the first kind of algorithms are k-means [Duda et al., 2001] or Self-Organizing Maps (SOM) [Kohonen, 1997], whereas the reader is referred to [Nister and Stewenius, 2006] or [Sun et al., 2010] for representative examples of the second. During the last few years other novel approaches have been proposed that consider the semantic embedding on visual vocabularies, so that the visual words are accurately selected by using semantic information from annotations [Ji et al., 2010]. In [van Gemert et al., 2010], the interested reader can find a comparison between several approaches for constructing visual vocabulary in a large-scale video retrieval scenario.

Generation of the BoW representation

Once a visual vocabulary has been constructed, the next step involves assigning each of the local descriptors in an image to the most similar word in the vocabulary and generate the corresponding image representation. Since the number of local descriptors in an image is variable (it depends on the detectors, the size of the image, and its content) and many machine learning techniques require fixed-length inputs, designing an input space that fulfills these requirements becomes a critical issue.

Assigning each local descriptor to a visual word in the vocabulary is basically a vector quantization process. Then, a histogram of word occurrence can be computed by counting the times that a visual word appears in an image. Of course, a normalization of the histogram by the total number of local regions is needed to provide comparable features for different images. This basic approach can be found in [Csurka et al., 2004].

However, a hard assignment of each sample to just one visual word does not take into account the similarity between visual words and, even more, does not consider the distance between a descriptor and its closest word. This may lead to performance loses in situations in which a descriptor is almost equally similar to two different visual words or a descriptor is not similar to any word in the vocabulary. To overcome this drawback some authors have proposed the use of soft-assignment in the histogram construction. This approach, found in [Philbin et al., 2008], computes a similarity measure with respect to all the words in the vocabulary and increments the histogram according to these similarities. The histogram is finally normalized by the total sum of the similarities.

Classification based on BoW representation

This section discusses different approaches for image classification based on the BoW representation. First of all, it is noteworthy how this problem can be seen as a binary or a multiclass categorization problem. Real world images hardly represent a unitary semantic concept but represent multiple objects and elements in a scene. Even when the concepts to be detected are very abstract or scene-oriented (such as indoor, outdoor, natural landscape, cityscape, ...) an image might be associated to more than one concept. For that reason, most challenges and standard evaluation approaches reformulate the image classification problem as a binary detection problem, thus involving the generation of several detectors (one for each concept) that label the images according to the presence or absence of a particular concept. This approach is followed by challenges like Pascal VOC [Everingham et al., 2009] or TRECVID Semantic Indexing task (formerly known as High Level Feature Extraction)[Smeaton et al., 2009]. On the the other hand, a good example of a benchmark that still considers image recognition as a multiclass problem is Caltech 256 [Griffin et al., 2007].

For image categorization, machine learning techniques are the most common approach, going from a simple Naive Bayes classifier [Lewis, 1998] to more advanced algorithms such as the Support Vector Machines (SVM) [Schölkopf and Smola, 2002], [Wallraven et al., 2003] [Grauman and Darrell, 2005]. In addition, generative frameworks [Hofmann, 2001] [Blei et al., 2003] also can make use of the BoW paradigm.

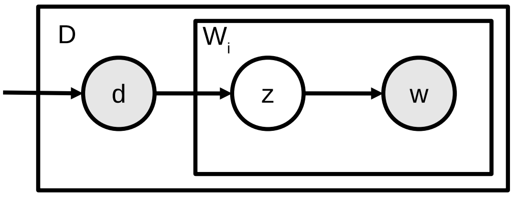


Figure 4.3: Graphical model representation of PLSA. Following the standard graphical model formalism, nodes represent random variables, while the edges show the dependencies among variables. Shaded circles denote observed variables and unshaded ones denote latent variables to be inferred. Boxes refer to different instances of the same variable.

4.2.2 Latent Topic Models

This section focuses on Latent Topic Models (LTMs). Though discriminative approaches currently show superior performance in some of the considered tasks, the potential of generative models can not be ignored since they provide underlying information that is not available in the discriminative framework. Furthermore, Latent Topic Models offer a fundamental advantage: they can be used in either unsupervised or supervised way.

The most common Latent Topic Models are the Probabilistic Latent Semantic Analysis (PLSA) [Hofmann, 2001] and Latent Dirichlet Allocation (LDA) [Blei et al., 2003]. Though both original PLSA and LDA are unsupervised algorithms, their formulation may be extended to handle different kinds of supervision (the interested reader is referred to [Blei and Mcauliffe, 2007] for a good example of supervised topic models), even partial supervision, where some labels are provided but others are missing [Ano, 2008]. Next subsections provide a brief review of these two fundamental LTMs to provide a proper background on which describing the original contributions of this work.

Probabilistic Latent Semantic Analysis

PLSA works on the classical image representation provided by the bag-of-words model. For each image, some potentially stable *keypoints* are detected and their corresponding descriptors are extracted from *local patches* around these keypoints. This kind of representation is no longer feasible when the task deals with a large number of images and a variable number of keypoints per image. In order to attain a more compact representation, the local patch descriptors are clustered around what are called *visual words*, to end up with an image represented as a *bag-of-visual-words* (see section 4.2.1). In the following, we will indistinctly refer to documents and images. Additionally, words and visual words will also be equivalent.

Fig. 4.3 illustrates the graphical model representation of PLSA. It relies on three variables: d represents the documents, w are the visual words that describe the appearance of the local patches, and z stand for the hidden topics hopefully related to semantic concepts, which explain the content of the images. From these variables, the underlying generative process modeled by PLSA represents the documents d as a mixture of latent topics z, which are supposed to be able to explain the occurrences of the visual words w in the documents.

Given a training corpus X consisting of D documents, and considering an M-word vocabulary and K latent topics, the corpus is summarized by means of an DxM co-occurrence table, denoted as N, where $n(d_i, w_j)$, with $i = 1, 2, ..., D$ and $j = 1, 2, ..., M$, represents the number of occurrences of the word w_j in the document d_i. The joint probability of documents and words can be obtained by marginalizing over the topics z_k with $k = 1...K$, as follows:

$$P(d_i, w_j) = P(d_i) \sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i) \tag{4.1}$$

where $P(d_i)$ is the probability that a word occurrence will be observed in a particular document d_i, $P(w_j|z_k)$ is the conditional probability of the word w_j given a particular topic z_k, and $P(z_k|d_i)$ is the conditional probability of the topic z_k given the document d_i.

In the training phase, the model parameters $\theta = \{P(w_j|z_k), P(z_k|d_i)\}$ are determined by maximizing the likelihood of generating the training corpus. The likelihood function accumulated over the training set obeys:

$$L = P(X|\theta) = \prod_{i=1}^{D} \prod_{j=1}^{M} P(d_i, w_j)^{n(d_i,w_j)} \tag{4.2}$$

where $P(d_i, w_j)$ is given by eq. (4.1). The Expectation-Maximization (EM) algorithm is used to obtain the optimal model parameters θ_{ML} by maximizing the log likelihood, i.e.:

$$\theta_{ML} = \arg\max_{\theta} \log P(X|\theta) \tag{4.3}$$

In the E-step, the posterior probabilities for the latent variables are computed from the current estimates of the parameters:

$$P(z_k|d_i, w_j) = \frac{P(z_k|d_i)P(w_j|z_k)}{\displaystyle\sum_{k=1}^{K} P(z_k|d_i)P(w_j|z_k)}, \tag{4.4}$$

and in the M-step, the re-estimation equations are used to update the parameter estimates:

$$P(z_k|d_i) = \frac{\displaystyle\sum_{j=1}^{M} n(d_i, w_j)P(z_k|d_i, w_j)}{n(d_i)} \tag{4.5}$$

$$P(w_j|z_k) = \frac{\displaystyle\sum_{i=1}^{D} n(d_i, w_j)P(z_k|d_i, w_j)}{\displaystyle\sum_{j=1}^{M}\sum_{i=1}^{D} n(d_i, w_j)P(z_k|d_i, w_j)} \tag{4.6}$$

where $n(d_i) = \sum_j n(d_i, w_j)$.

The interested reader is referred to [Hofmann, 2001] for a comprehensive explanation of the model and a complete derivation of the previous equations.

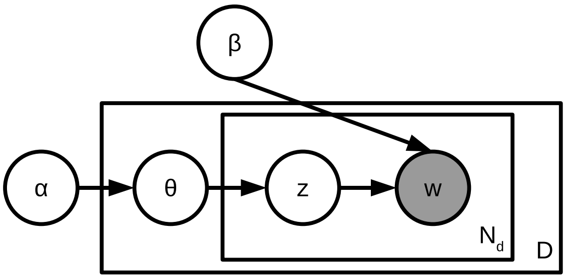


Figure 4.4: Graphical model representation of LDA. Following the standard graphical model formalism, nodes represent random variables, while the edges show the dependencies among variables. Shaded circles denote observed variables and unshaded ones denote latent variables to be inferred. Boxes refer to different instances of the same variable.

Latent Dirichlet Allocation

The graphical model of LDA is shown in Figure 4.4. It is noteworthy that LDA uses a different notation than the PLSA. In order to properly explain the generative process defined by LDA, we first describe the observable variables involved in the model. Given a collection D of images (corpus), each image $d \in D$ is described by means of a set of N_d local patches $n \in N_d$, each of them being indexed by a visual descriptor w_n that describes its appearance.

Intuitively, the whole corpus is modeled by a parameter $\boldsymbol{\alpha}$ that sets the global distribution of the topics in the corpus (topic proportions). Then, for each document d, a new variable $\boldsymbol{\theta}_d$ stores the particular distribution of topics in the document. This new variable is used to choose the topic z_n associated to each local patch n in the image so that, depending on that topic, the appearance (visual descriptor) associated to the local patch w_n is assigned. Since each topic tends to generate particular visual descriptors, the final content of a document depends on the topics it contains and their proportions.

From this explanation, it is easy to notice that the main difference between LDA and PLSA lies in the fact that former additionally learns the prior distributions of the topics in the corpus $\boldsymbol{\alpha}$ so that it can make use of this information when the model is used on new sets of unlabeled images (thus avoiding overfitting). Consequently, the generative process of LDA is as follows:

1. For each document d, sample a Dirichlet random variable $\boldsymbol{\theta}|\boldsymbol{\alpha} \sim Dir(\boldsymbol{\alpha})$ that provides a probability distribution over the K latent topics.
2. For each local patch n, $n \in N$:
   (a) Sample a topic $z_n|\boldsymbol{\theta} \sim Mult(\boldsymbol{\theta})$.
   (b) Draw its appearance as $w_n|z_n, \boldsymbol{\beta} \sim Mult(\boldsymbol{\beta}_{z_n})$

where $Mult(\cdot)$ stands for a multinomial distribution.

For the sake of compactness, we omit the subindex d in those variables that are document-dependent unless a sum over the documents is performed. Considering a particular document d in the corpus, several parameters are involved in its generative process:

- $\boldsymbol{\alpha}$ is a K-dimensional vector that contains the parameters $\alpha_k > 0$ of the Dirichlet distribution. This parameter is shared by all the documents in the corpus.
- $\boldsymbol{\beta}$ is a collection of K V-dimensional vectors $\boldsymbol{\beta}_k = [\beta_{k1} \dots \beta_{kV}]$ containing the probabilities of the visual words given the latent topics.

For each document in the corpus, the resulting joint distribution on visual words, spatial locations, and hidden variables is given by:

$$p(\mathbf{w}, \boldsymbol{\theta}, \mathbf{z}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \prod_{n=1}^{N_d} p(z_n|\boldsymbol{\theta}) p(w_n|z_n, \boldsymbol{\beta}) \tag{4.7}$$

The key inferential problem of LDA is that of computing the posterior distribution of the hidden variables given a document:

$$p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{p(\mathbf{w}, \boldsymbol{\theta}, \mathbf{z}|\boldsymbol{\alpha}, \boldsymbol{\beta})}{p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})} \tag{4.8}$$

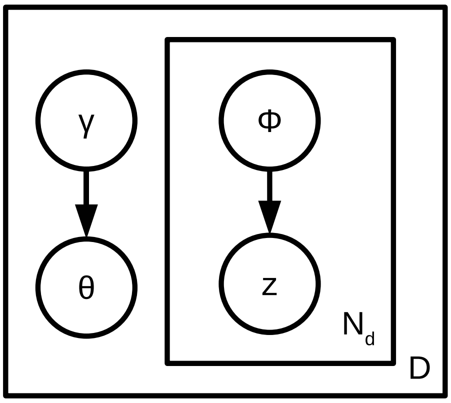


Figure 4.5: Graphical model of the variational distribution used to approximate the posterior in LDA. Following the standard graphical model formalism, nodes represent random variables, while the edges show the dependencies among variables. Shaded circles denote observed variables and unshaded ones denote latent variables to be inferred. Boxes refer to different instances of the same variable.

Unfortunately, this distribution is intractable due to the coupling between $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Thus, the authors [Blei et al., 2003] propose the use of mean-field variational methods for approximate inference [Jordan et al., 1999]. The basic idea of the proposed method is to make use of Jensen's inequality to obtain an adjustable lower bound on the log likelihood that is indexed by several variational parameters. Then the values of the variational parameters are optimized as an attempt to find the tightest possible lower bound.

In particular, the authors proposed to use a simplified graphical model in which some of the nodes and edges were removed. Figure 4.5 shows the selected variational model for LDA, which gives place to a variational distribution q that follows:

$$q(\theta, \mathbf{z}|\boldsymbol{\gamma}, \boldsymbol{\phi}) = q(\theta|\gamma) \prod_{n=1}^{N_d} q(z_n|\phi_n) \tag{4.9}$$

where the Dirichlet parameter γ and the multinomial parameters $(\phi_1, ..., \phi_{N_d})$ are the free variational parameters. This variational distribution shows independence between the variables, what makes it tractable. Finally, as described in the original paper [Blei et al., 2003], providing the tightest lower bound of the posterior translates to the following optimization problem:

$$(\boldsymbol{\gamma}^*, \boldsymbol{\phi}^*) = \arg\min_{(\boldsymbol{\gamma},\boldsymbol{\phi})} D(q(\boldsymbol{\theta}, \mathbf{z})|p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})) \tag{4.10}$$

where D stands for the KL divergence. Hence, the inference process requires to minimize the Kullback-Leibler (KL) divergence between the true posterior and a variational distribution, which represents a simplified but convenient version of the original model.

Modeling the spatial distribution of visual words in LTMs

Undoubtedly, the most important limitation of PLSA and LDA in Computer Vision is that they do not take into account the spatial distribution of visual words in the images. The benefits of this spatial modeling are twofold: first, an improved performance of Latent Topic Models in tasks such as image classification or topic discovery; and second, an enrichment of such models with the capability of generating robust image segmentations. However, modeling the spatial location of visual words under this framework is not longer straightforward since one must ensure that both appearance and spatial models are jointly trained using the same learning algorithm that infers the latent topics. One of the first approaches considering some geometric modeling was [Sivic et al., 2005], where the use of doublets of visual words over PLSA added simple geometric considerations and achieved notable improvements in object localization. In [Fergus et al., 2005] the authors modeled the joint distribution of visual features and their locations using a translation and scale invariant approach for unsupervised category discovering. This proposal assumes that the objects tend to consistently appear at a predefined set of spatial positions and scales or, at least, the objects are large enough to "fit" in those locations. In [Liu and Chen, 2006], Gaussian and uniform spatial distributions are used to model foreground and background topics, respectively, thus providing simple shape and scale estimates of the objects to be discovered and classified. However, a Gaussian distribution still represents a coarse approximation of the object shape; thus, although the model improves the object localization, it is not able to produce high-quality segmentations. A similar idea is explored in [Wang and Grimson, 2007], where LDA is extended so that the documents are no longer images, but points in images; then, a new latent variable associates visual words with documents and the spatial location is modeled as a Gaussian centered at the document. In [Bosch et al., 2008] PLSA is extended by including a Spatial Pyramid to come up with what they call the SP-PLSA (Spatial Pyramid-Probabilistic Latent Semantic Analysis). Other kind of approaches encode geometry information using what is known as a part model, in which the objects are composed of parts that, in turn, are shared among different categories. The work in [Sudderth et al., 2007] represents a good example of a hierarchical part-based model using latent topics.

Other proposals take a step forward and incorporate previous blind segmentations of images into the Latent Topic Models. In [Zhang and Zhang, 2004], a version of PLSA that considers topics at region level (from a previous segmentation) is proposed for image retrieval. In [Russell et al., 2006], a novel approach to deal with under- and over-segmentations is proposed. Multilevel segmentations are generated, then PLSA is used to unsupervisely detect categories, and finally the best segmentation level is chosen according to the distance between the proposed regions and the detected categories. In [Cao and Fei-Fei, 2007], an extension of LDA is proposed that considers topics at an intermediate level (regions). These topics produce two kinds of visual words, one related to the color of the whole region, and the other associated with texture descriptors from the local patches within the region. Thus, the algorithm starts from an over-segmented version of the image to end up with a more realistic segmentation, where regions are (hopefully) associated with semantic concepts. Similar approaches have been successfully applied to image classification and annotation [Wang et al., 2009], as well as to scene understanding [Li et al., 2009], where concurrent image annotation and segmentation as well as scene classification are achieved within an integrated framework. In these approaches, the method for image annotation follows the model proposed in [Blei and Jordan, 2003]. In particular, annotations can be seen as image captions that have been generated by specific latent topics. Obviously, these image captions may help to classify the concept represented by the whole scene; however, since they do not point at any specific region, the association of captions with regions is made through latent variables that need to be inferred during the training phase. In all the reviewed models, however, regions are considered as independent entities that do not interact with each other. One method that goes beyond and, by allowing interactions among regions, imposes certain spatial coherence has been proposed in [Zhao et al., 2010]. In this work, a Markov Random Field (MRF) enforces that spatially connected regions belong to the same topics.

The experimental results in all these previous works support the idea of modeling the spatial distribution of visual words as a promising way to improve current latent topic model performance.

4.3 Region-Based Latent Topic Model

The Region-Based Latent Topic Model (RBLTM), proposed in [González-Díaz et al., 2009a], aims to incorporate the spatial location of visual words into a latent topic model. In particular, RBLTM extends the PLSA model and considers a document as a set of inter-related regions that influence to each other according to their closeness. Two are the main contributions of this model with respect to the state-of-the-art in Latent Topic Models: 1) RBLTM considers regions as active entities, which can interact with each other rather than simply representing concepts by their appearance; and 2) it provides a formal framework for supervised training. Two kinds of annotations may be considered for supervised training: image-based weak annotations, such as image captions, and (semi)strong region-based annotations. The latter allows us to enhance the classification performance when bounding-box or pixel-wise ground truth segmentations are available. Both contributions come up from the manner in which RBLTM models the spatial distribution of topics over the document, which is novel with respect to previous related works [Wang et al., 2009][Blei and Jordan, 2003][Li et al., 2009]. Furthermore, strong and weak levels of supervision can coexist in our model, as it will be shown in our experiments, that demonstrate that RBLTM successfully addresses three tasks of interest in Computer Vision: image classification, object class image segmentation and unsupervised topic discovery.

As previously mentioned, the original PLSA does not take into account any spatial information. As a result of this lack of constraints on the spatial position of the words, the distribution of topics over the visual words often turns out inaccurate. Consequently, the topic detection process does not work properly, and the modeling of semantic concepts becomes infeasible. In contrast, the goal of RBLTM is to exploit both the local descriptors, which have been shown to be highly discriminative, and the image segmentation, which depicts the spatial structure.

In this Section, the preprocessing stage devoted to set up the bag-of-words model on top of which RBLTM is built is described first. Next, the unsupervised version of RBLTM is explained, introducing the basic formulation of the model. Subsequently, the supervised version of the model, which allows for improved operation when annotated datasets are available, is presented. Finally, a brief discussion about the computational complexity of RBLTM is provided.

4.3.1 Preprocessing: setting up the bag-of-words model

Since the proposed scheme makes use of both local properties and global segmentation-based information, the images have to be preprocessed in order to extract and properly organize the required information for the subsequent model learning process. Fig. 4.6 illustrates the image representation used in the proposed scheme. Each image i is partitioned into R_i regions, and a matrix Λ^i is defined, as described below, to model the inter-region influences. Then, keypoints are detected and the corresponding W_i local patches are described through both their appearance (color and texture) w_j and their spatial location s_l^i, which links each keypoint to the specific region to which it belongs.

The preprocessing module involves several stages, namely: generation of the image segmentation, computation of the inter-region influence matrix, extraction of local features, and generation of the visual vocabulary. Each of these stages is briefly described in the next paragraphs.

a) Image segmentation: The segmentation stage uses a fast algorithm that is particularly configured to produce about 30-60 regions (i.e., an over-segmentation). This configuration ensures that the regions usually contain pixels from only one semantic object. Specifically, an efficient graph-based image segmentation method [Felzenszwalb and Huttenlocher, 2004] is employed to generate color-based segmentations for each image. We have employed this method because our experiments, presented in chapter 5, use a still-image database rather than a video database. In the second case, our blind segmentation proposal described in chapter 2 would provide even better results.

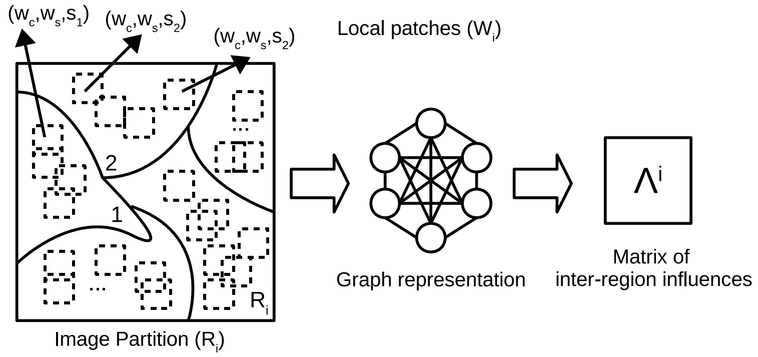


Figure 4.6: Image representation used in the RBLTM. Each image i is partitioned into R_i regions that are organized in a graph whose nodes represent regions and whose links stand for influences among regions. This graph representation sets the basis to compute a matrix of inter-region influences. On the other hand, keypoints are detected and the corresponding W_i local patches (dashed boxes) are described through both their local appearance (color w_c and texture (SIFT) w_s visual words) and their spatial location (s_l, with l the region that contains the patch).

b) Inter-region influence matrix: The proposed generative model uses a matrix $\Lambda^i := \left(\lambda^i_{pl}\right)_{R_i R_i}$ that holds information about the relations among regions that come from the segmentation stage. The relations among regions are modeled through a simple influence model. Specifically, given an image i partitioned into a set of R_i regions, the influence λ^i_{pr} of a region p on a region r is measured as follows:

$$\lambda^i_{pr} = \frac{l_{pr}}{l_r} \tag{4.11}$$

where l_{pr} is the length of the common boundary between the regions p and r, and l_r represents the perimeter length of the region r. Values close to 1 mean that the influence of p on r is strong and vice versa. It is worth noting that the matrix Λ^i is not, in general, symmetric, i.e., $\lambda^i_{pr} \neq \lambda^i_{rp}$. Then, the influence of larger regions on smaller ones is higher than in the opposite way, what, from the authors' point of view, is desirable in order to avoid that small regions with just a few visual words (thus, with low confidence on their class) present a strong influence over large regions with many visual words. Furthermore, as one can expect, the influence of a region on itself is $\lambda^i_{pp} = 1$.

c) Local feature extraction: Local features are extracted from each local patch in every image. The keypoints and their corresponding local patches are obtained using a dense grid. In particular, two scales have been used, yielding overlapped circular patches with radius 8, and 16, organized in a regular 6 pixel spaced grid. Hence,for each selected location in the image, two independent local patches are generated, one associated to each of the scales. From each local patch, two kinds of appearance features are extracted. The first one, w_c, is related to color and consists of a 96-dimensional vector formed by the concatenation of 4 spatial histograms (2x2 grid)of 24 color components in the CIELab space; and the second one, w_s, consists of an 128-dimensional SIFT descriptor [Lowe, 2004] that models texture information.

d) Visual Vocabulary: Once the local features have been extracted, a bag-of-words model is computed. The k-means clustering algorithm has been used to compute the M codewords that best represent the local features of the reference image set. In our case, given a complete set of $1M$ descriptors, the k-means algorithm provided vocabularies of size $M_s = 4000$ for SIFT features and $M_c = 1000$ for color features. Moreover, the influence of each type of features is adjusted by means of regularizing priors. The relative weight of color features was set to 0.5 by cross-validation, and regularization is used according to this weight. For simplicity, in the remainder of this section, although we use two kinds of visual words, we will refer to a general visual word w. The extension of the formulation to more than one visual word is straightforward, since their distributions are conditionally independent. Hence, they can be factorized and the corresponding equations adopt the same form.

4.3.2 Unsupervised RBLTM

The graphical model representation of RBLTM is shown in Fig. 4.7. A new variable s has been introduced to indicate the spatial location to which a local descriptor refers; this new variable becomes the most significant difference with respect to PLSA, which

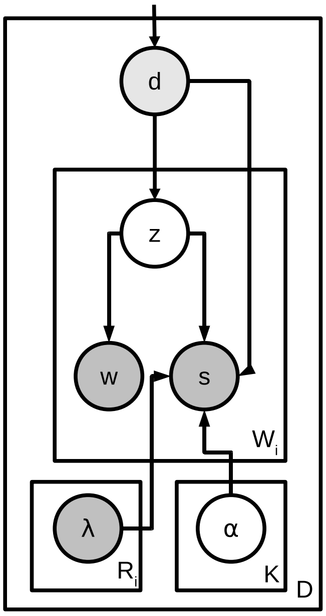


Figure 4.7: Graphical model representation of unsupervised RBLTM. Following the standard graphical model formalism, nodes represent random variables, while the edges show the dependencies among variables. Shaded circles denote observed variables and unshaded ones denote latent variables to be inferred. Boxes refer to different instances of the same variable.

does not take into consideration the spatial layout of the image. The objective of the spatial modeling is to improve the decisions concerning those regions that cannot be consistently characterized by their appearance, by taking also into consideration some information from their neighborhood. In particular, the spatial location of a visual descriptor is given by the region to which the local patch belongs (see Figure 4.6). Consequently, for each image i there are as many potential locations as regions R_i have been generated in the segmentation stage.

The other new variable in the graph, $\boldsymbol{\alpha}$, will be defined later on when the basic model formulation has been introduced.

As illustrated in the graph, the generative process modeled by RBLTM is as follows. Each document d_i is represented as a mixture of latent topics z_k, $k =$

$1, 2, ..., K$ that are supposed to explain the occurrences of visual words w_j, $j = 1, 2, ..., M$ at a predetermined set of spatial locations s_l^i, $l = 1, 2, ..., R_i$, which are document-dependent. The visual words w_j describe the local appearance of the local patches (texture and color), while the spatial locations s_l^i denote to which R_i region the local patch belongs. Since this notation is used in the rest of the section, it should be noted that, for any collection of random variables, the subscripts denote the working index of a variable itself, while the superscripts denote the indices along the collection.

In RBLTM, each document in the corpus is described by a co-occurrence MxR_i table N^i with $n^i(w_j, s_l^i)$, whose dimensions differ from one image to another; consequently, a global co-occurrence table makes no sense in this case. Following the same steps developed for PLSA, the joint distribution of documents, topics, and words can be written as follows by marginalizing over the topics:

$$P(d_i, w_j, s_l^i) = P(d_i) \sum_{k=1}^{K} P(z_k|d_i) P(w_j|z_k) P(s_l^i|z_k, d_i, \boldsymbol{\alpha}) \tag{4.12}$$

In this case, the likelihood function to optimize becomes:

$$L = P(X|\theta) = \prod_{i=1}^{D} \prod_{j=1}^{M} \prod_{l=1}^{R_i} P(d_i, w_j, s_l^i)^{n^i(w_j, s_l^i)} \tag{4.13}$$

and the optimal model parameters θ_{ML} are again found by maximizing the accumulated log likelihood, i.e.:

$$\theta_{ML} = \arg\max_{\theta} \log P(X|\theta) \tag{4.14}$$

by means of the EM algorithm.

In order to perform the log likelihood optimization, the conditional probability $P(s_l^i|z_k, d_i, \boldsymbol{\alpha})$ has been modeled as a parametric distribution of the form:

$$P(s_l^i|z_k, d_i, \boldsymbol{\alpha}) = \sum_{p=1}^{R_i} \alpha_p^{ik} \lambda_{pl}^i \tag{4.15}$$

where $\boldsymbol{\alpha}$ is a collection of KxD unknown parameters $\boldsymbol{\alpha}^{ik} = [\alpha_1^{ik} \dots \alpha_{R_i}^{ik}]$ that has to inferred during the learning phase. Each element of the vector is called the *importance* α_p^{ik} of a region p given a topic k. The dependence of the distribution on z_k and d_i is trivial since they basically point to the corresponding element $\boldsymbol{\alpha}^{ik}$ in the collection. The term λ_{pl}^i was already defined in eq. (4.11) as the *influence* of the region s_p^i on the region s_l^i – the influences are computed "a priori" for each image in the dataset as described in subsection 4.3.1. The role of the influences is to induce the topics to spread over contiguous regions; consequently producing spatially coherent and compact topics. In other words, when the likelihood of topic is high given a region, this region will shift its spatially neighboring regions towards that topic. An example of the spatial distribution for a particular image is presented in Figure 4.8. This example shows the empirical spatial distributions for topics representing the classes "dog" and "sheep" at several iterations of the algorithm. The figure demonstrates how the algorithm converges to very close representations of the objects (it is noteworthy that this spatial distribution does not explicitly take into account the appearance of the objects). Furthermore, last column shows the results of the RBLTM when there is no cooperation among regions: in this case, there are several regions that are not in agreement with their spatial neighborhood, issue that is successfully handled by the cooperative model. Hence, we can conclude that this modeling is simple, allows us to obtain a closed solution that enables a joint optimization of the appearance and spatial models, and provides excellent results as it will be shown in chapter 5.

Concerning the maximization of the log-likelihood, in the E-step of the EM algorithm the posterior probabilities for the concepts $P(z_k|d_i, w_j, s_l^i)$ are computed from the current estimates of the parameters as follows:

$$P(z_k|d_i, w_j, s_l^i) = \frac{P(z_k|d_i)P(w_j|z_k)P(s_l^i|z_k, d_i, \boldsymbol{\alpha})}{\sum_{m=1}^{K} P(z_m|d_i)P(w_j|z_m)P(s_l^i|z_m, d_i, \boldsymbol{\alpha})} \tag{4.16}$$

(The derivations of this formula and those that follows are given in the Appendix

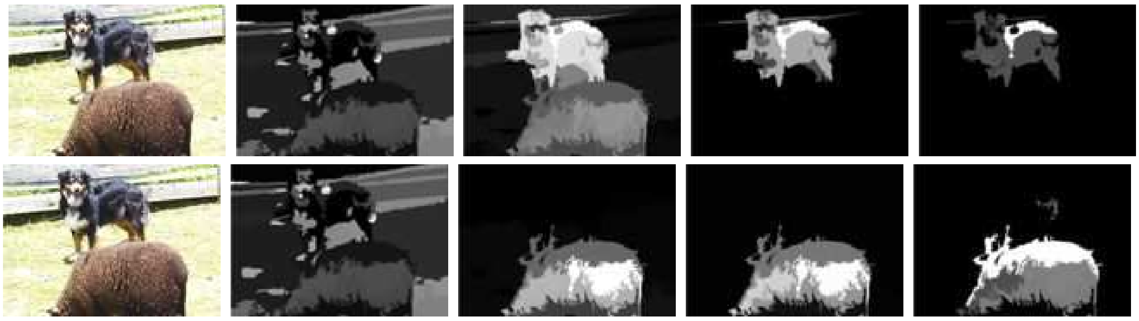


Figure 4.8: An example of the spatial distribution for classes "dog" (top row) and "sheep" (bottom row) in a image of the database: First column: original images; Second column: Iteration 1; Third column: Iteration 10; Fourth column: Iteration 30; Fifth column: Iteration 30 without cooperation (a region just influences itself). Lighter colors represent higher probabilities.

B.1.)

In the M-step, $P(z_k|d_i)$ and $P(w_j|z_k)$ are re-estimated as:

$$P(z_k|d_i) = \frac{\sum_{j=1}^{M}\sum_{l=1}^{R_i} n^i(w_j, s_l^i) P(z_k|d_i, w_j, s_l^i)}{\sum_{j=1}^{M}\sum_{l=1}^{R_i} n^i(w_j, s_l^i)} \tag{4.17}$$

$$P(w_j|z_k) = \frac{\sum_{i=1}^{D}\sum_{l=1}^{R_i} n^i(w_j, s_l^i) P(z_k|d_i, w_j, s_l^i)}{\sum_{i=1}^{D}\sum_{l=1}^{R_i} n^i(w_j, s_l^i)} \tag{4.18}$$

Additionally, $P(s_l^i|z_k, d_i, \boldsymbol{\alpha})$ is re-estimated following eq. (4.15) and using updated estimates of the importances, which are computed as follows:

$$\alpha_p^{ik} = \frac{\sum_{j=1}^{M}\sum_{l=1}^{R_i} n^i(w_j, s_l^i) P(z_k|d_i, w_j, s_l^i) r_{pl}^{ik}}{\chi_p^i \cdot \sum_{j=1}^{M}\sum_{l=1}^{R_i} n^i(w_j, s_l^i) P(z_k|d_i, w_j, s_l^i)} \tag{4.19}$$

where the vector $\boldsymbol{\chi}^i = \left[\chi_1^i \dots \chi_{R_i}^i\right]$ contains the weighting factors χ_p^i that accumulate the total influence of a particular region over the rest (see Appendix B.1.3):

$$\chi_p^i = \sum_{l=1}^{R_i} \lambda_{pl}^i, \tag{4.20}$$

and r_{pl}^{ik} is a normalized factor (it satisfies $\sum_{l=1}^{R_i} r_{pl}^{ik} = 1$) that considers the whole (importance plus influence) inter-region relations given a topic, and it is defined as follows (see Appendix B.1.2):

$$r_{pl}^{ik} = \frac{\alpha_p^{ik} \lambda_{pl}^i}{\displaystyle\sum_{m=1}^{R_i} \alpha_m^{ik} \lambda_{ml}^i}. \tag{4.21}$$

The computation of r_{pl}^{ik} is performed in the E-step of the algorithm, once the updated estimates of the importances are available.

The EM algorithm to perform the complete inference process in the unsupervised RBLTM is summarized in Alg. 1.

4.3.3 Supervised RBLTM

RBLTM can be modified to work in a supervised framework, in which a set of annotated images is available for the training phase. Furthermore, the proposed supervised extension considers two possibilities for the annotations concerning the spatial structure of a document: image-based and region-based annotations. A graphical model of the supervised version of RBLTM is shown in Fig. 4.9. As it can be seen in the figure, the graph shows two new parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ that depend on the image L_{IMG} and region L_{REG} based annotations, respectively. Given a document i, $\boldsymbol{\beta}^i$ is a K-dimensional multinomial variable that contains the proportions of the topics, whereas $\boldsymbol{\gamma}^i$ is a collection of K R_i-dimensional multinomial variables $\boldsymbol{\gamma}^{ik}$ that define the spatial location of a topic in the document. In practice, $\boldsymbol{\beta}^i$ and $\boldsymbol{\gamma}^{ik}$ are the parameters of two Dirichlet distributions that become the priors of $p(z_k|d_i)$ and the importances $\boldsymbol{\alpha}^{ik}$, respectively.

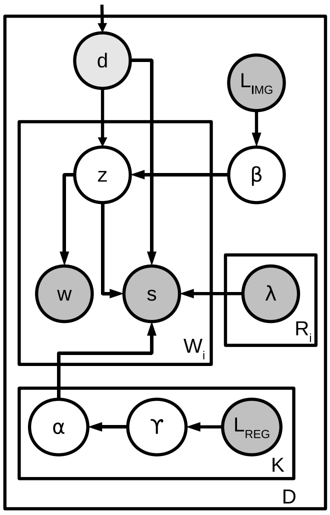


Figure 4.9: Graphical model representation of supervised RBLTM with strong and weak annotations. Following the standard graphical model formalism, nodes represent random variables, while the edges show the dependencies among variables. Shaded circles denote observed variables and unshaded ones denote latent variables to be inferred. Boxes refer to different instances of the same variable.

In general, the inclusion of prior distributions leads to a Maximum a Posteriori (MAP) optimization over the parameters θ_{MAP}:

$$\theta_{MAP} = \arg\max_{\theta} \{\log P(X|\theta) + \log g(\theta)\} \tag{4.22}$$

where $g(\theta)$ stands for the prior density of the parameters. This prior density function $g(\theta)$ can be factorized as $g(\theta) = g_{img}(\theta) \cdot g_{reg}(\theta)$, which are related to the image-based and the region-based annotations, respectively. In both cases conjugate priors represent good candidates for Bayesian inference, which leads to the use of Dirichlet priors for the multinomial distributions. Hence, our proposal in the supervised scenario uses soft-labeling in the sense that topics are still latent. In practice, a document or region labels do not impose that every local patch in the image/region has to be assigned to the topic associated to the label (e.g. if the appearance of the topic does not fit the learned distribution it may be assigned to other topic).

In the next two subsections the equations that differs with respect to those of the unsupervised case are given for both image- and region-based annotations, respectively. The complete derivation of these equations is provided in the Appendix B.2. The rest of the equations are the same as the corresponding ones in the unsupervised version of the model.

Image-based annotations

An image is labeled as a positive example when it contains an object of interest. In this case, a Dirichlet prior over the distribution of the topics given the document is considered. Specifically, the prior density function obeys:

$$g_{img}(\theta) = \prod_{i=1}^{D} \frac{1}{G(\beta^i)} \prod_{k=1}^{K} P(z_k|d_i)^{(\beta_k^i - 1)} \tag{4.23}$$

where β_k^i represent the hyperparameters of the Dirichlet densities, and the normalizing constant G follows:

$$G(\beta^i) = \frac{\prod_{k=1}^{K} \Gamma(\beta_k^i)}{\Gamma\left(\sum_{k=1}^{K} \beta_k^i\right)} \tag{4.24}$$

Then, as a result of a Lagrangian optimization subject to $\sum_k P(z_k|d_i) = 1$, the formula to update the MAP estimates of the $P(z_k|d_i)$ in the M-step of the EM algorithm turns out to be as follows:

$$P(z_k|d_i) = \frac{\displaystyle\sum_{j,l} n^i(w_j, s_l^i) P(z_k|d_i, w_j, s_l^i) + (\beta_k^i - 1)}{\displaystyle\sum_{j,l} n^i(w_j, s_l^i) + \sum_{m} (\beta_m^i - 1)} \tag{4.25}$$

where $j = 1, 2, ..., M$, $l = 1, 2, ..., R_i$ and $m = 1, 2, ..., K$. The complete derivation of this formula is worked out in the Appendix B.2.2.

Furthermore, the hyperparameters can be modeled as $\beta_k^i = \epsilon_{IMG}^i \tilde{P}(z_k|d_i) + 1$, where ϵ_{IMG}^i is shared by all the topics of a document, and $\tilde{P}(z_k|d_i)$ stands for the prior probability of the topic k given the image i. These prior probabilities can be easily determined using the image-based labels, henceforth L_{IMG}. Hence, the eq. (4.25) can be rewritten as follows:

$$P(z_k|d_i) = \frac{\sum_{j,l} n^i(w_j, s_l^i) P(z_k|d_i, w_j, s_l^i) + \epsilon_{IMG}^i \tilde{P}(z_k|d_i)}{\sum_{j,l} n^i(w_j, s_l^i) + \epsilon_{IMG}^i} \tag{4.26}$$

where $j = 1, 2, ..., M$ and $l = 1, 2, ..., R_i$. As it can be easily inferred from the update expression, the term ϵ_{IMG}^i manages the balance between the contributions of the unknown and known terms of the distribution.

Region-based annotations

The proposed model also considers a simple and effective way to use region-based labels during the training phase. The region-based annotations, henceforth L_{REG}, refer to either pixel-wise image segmentations, when every pixel is labeled as belonging to a specific class (object), or bounding-box-based annotations, when a rectangular frame is given that contains the object of interest. In the second case, although only partial information is available (just a rectangular area) and not all the pixels are correctly labeled, a mask associated to each of the objects happening in the image can be generated.

In this case, a conjugate prior is set over the *importance* α_p^{ik} of each region given the topic. The importance vector $\boldsymbol{\alpha}^{ik}$ could be considered as a multinomial variable, except for the normalizing factors χ_p^i. Thus, considering the Hadamard product $\left(\boldsymbol{\alpha}^{ik} \cdot \boldsymbol{\chi}^i\right)$ as a multinomial variable, a Dirichlet prior can be written as follows:

$$g_{reg}(\theta) = \prod_{i=1}^{D} \prod_{k=1}^{K} \frac{1}{G(\gamma^{ik})} \prod_{p=1}^{R_i} (\alpha_p^{ik} \chi_p^i)^{(\gamma_p^{ik}-1)} \tag{4.27}$$

where γ_p^{ik} are the hyperparameters of the distribution, and the normalizing constant G follows:

$$G(\gamma^{ik}) = \frac{\prod_{p=1}^{R} \Gamma(\gamma_p^{ik})}{\Gamma\left(\sum_{p=1}^{R} \gamma_p^{ik}\right)} \tag{4.28}$$

Furthermore, we have designed the prior distribution by introducing some importance priors $\tilde{\alpha}_p^{ik}$ as follows: once the masks for the different objects have been computed (either from pixel-wise segmentations or from bounding boxes), a mapping between them and the segmentation described in Subsection 4.3.1 is performed. Such a correspondence is not previously available since, in general, the segmentation of the preprocessing stage provides many more regions than the ground truth solutions (ground truth models semantic concepts, which are usually compound by more than one homogeneous regions). In this work, a straightforward method has been designed to establish that mapping between each mask of a class c and a region p of the partition $\{R_i\}$. The underlying idea is to compute the ratio between the number of pixels of the region that belongs to a specific class and the total number of pixels of that region. Let $l^i(x, y)$ denote the class label of the pixel (x, y) in an image i, $N_p^i = \{(x, y) \in p\}$ the total number of pixels of the region p, and $N_{pc}^i = \{(x, y) \in p | (l(x, y) = c)\}$ the number of pixels of the region p that belongs to the class c. Hence, the importance priors obey $\tilde{\alpha}_p^{ik} = \frac{|N_{pc}^i|}{|N_p^i|}$, with $k = c$.

Then, as in the previous cases, the Lagrangian optimization (worked out in the Appendix B.2.3) yields the following update equation for the *importances*:

$$\alpha_p^{ik} = \frac{\displaystyle\sum_{j,l} n^i(w_j, s_l^i) P(z_k | d_i, w_j, s_l^i) r_{pl}^{ik} + \epsilon_{REG}^{ik} \tilde{\alpha}_p^{ik}}{\chi_p^i \left[\displaystyle\sum_{j,l} n^i(w_j, s_l^i) P(z_k | d_i, w_j, s_l^i) + \epsilon_{REG}^{ik} \tilde{\Gamma}^{ik} \right]} \tag{4.29}$$

It should be noted that, once more, the hyperparameter γ_p^{ik} has been modeled as $\gamma_p^{ik} = \epsilon_{REG}^{ik} \tilde{\alpha}_p^{ik} + 1$, where $\tilde{\alpha}_p^{ik}$ are the priors of the importances. Again, the parameter ϵ_{REG}^{ik} manages the balance between the contributions of the unknown and known elements of the distribution. The term $\tilde{\Gamma}^{ik}$ represents a scaling factor that ensures that the spatial distribution is multinomial, and is defined as:

$$\tilde{\Gamma}^{ik} = \sum_p \tilde{\alpha}_p^{ik} \tag{4.30}$$

It is worth noting that this kind of annotation is not possible in PLSA, since it is not able to model the spatial distribution of topics in an image.

The Alg. 2 summarizes the complete inference process for the supervised RBLTM, including both the image- and the region-based labels.

4.3.4 A comment on the RBLTM complexity

RBLTM improves the image representation by PLSA through the modeling of the spatial location of latent topics and providing a natural framework for inter-region relations. However, this improvement is achieved in exchange for an increase of the computational complexity of the EM algorithm. In order to get some insight on this issue, let us contemplate a simplified model that considers K topics in a corpus of D documents, each of them containing the same number of local patches W and regions R. In this case, the computation of the terms related to the spatial location of topics requires $O(DKR^2)$ operations, that differs from the maximal complexity of the terms that are also present in the PLSA (those ones that do not include any spatial information), which is $O(DWK)$. Though the quadratic dependence on R might potentially cause some problems for high values of R, in practice, the complexity is limited by the model of the inter-region influences. In particular λ^i_{pl} is equal to zero for any pair of non-adjoining regions (see equation (4.11)); in other words, each region affects only to a few regions in its spatial neighborhood, what leads to an actual complexity of $O(DKnR)$ operations, with $n \approx 5$ in our experiments.

Algorithm 1 EM algorithm to perform the complete inference process in the unsupervised RBLTM

```
randomly initialize all the variables
repeat
  E-Step
  for all documents $d_i, i \in D$, visual words $w_j, j \in M$, and spatial locations $s^i_l, l \in R_i$ do
    compute $P(z_k|d_i, w_j, s^i_l)$, in eq. (4.16).
    for all regions $p \in R_i$ do
      compute $r^{ik}_{pl}$, in eq. (4.21).
    end for
  end for
  M-Step
  for all documents $d_i, i \in D$, visual words $w_j, j \in M$, spatial locations $s^i_l, l \in R_i$, and regions $p \in R_i$ do
    compute $P(z_k|d_i)$, in eq. (4.17).
    compute $P(w_j|z_k)$, in eq. (4.18).
    compute $\alpha^{ik}_p$, in eq. (4.19).
    compute $P(s^i_l|z_k, d_i, \boldsymbol{\alpha})$, in eq. (4.15)
  end for
until convergence
```

Algorithm 2 EM algorithm to perform the complete inference process in the supervised RBLTM

```
initialize $\tilde{P}(z_k|d_i)$ from the image-based labels $L_{IMG}$.
initialize $\tilde{\alpha}_p^{ik}$ from the region-based labels $L_{REG}$.
randomly initialize the rest of the variables
repeat
  E-Step
  for all documents $d_i, i \in D$, visual words $w_j, j \in M$, and spatial locations $s_l^i, l \in R_i$ do
    compute $P(z_k|d_i, w_j, s_l^i)$, in eq. (4.16).
    for all regions $p \in R_i$ do
      compute $r_{pl}^{ik}$, in eq. (4.21).
    end for
  end for
  M-Step
  for all documents $d_i, i \in D$ , visual words $w_j, j \in M$, and spatial locations $s_l^i, l \in R_i$ and $p \in R_i$ do
    compute $P(z_k|d_i)$, in eq. (4.26).
    compute $P(w_j|z_k)$, in eq. (4.18).
    compute $\alpha_p^{ik}$, in eq. (4.29).
    compute $P(s_l^i|z_k, d_i, \boldsymbol{\alpha})$, in eq. (4.15).
  end for
until convergence
```

4.4 Region Based Latent Dirichlet Allocation

Region-Based LDA (RBLDA) constitutes an advanced latent topic model that uses successful insights from RBLTM as well as other novel extensions and is developed under the LDA paradigm. As mentioned above, LDA learns corpus-level priors over the topic distribution that are not considered in PLSA. This enhancement motivates the task of adapting the original RBLTM to this more Bayesian framework. Furthermore, RBLDA has been conceived to overcome several particular drawbacks of the RBLTM, namely:

- RBLTM considers the appearance probabilities $p(w|z)$ as conditionally independent given the topic. Hence, it does not explore the nonlinear relations among words that belong to the same region.

- RBLTM considers simple intra-topic inter-region influences. This approach is somewhat constraining since it cannot model the influence between neighboring regions that belong to correlated topics (sky/aeroplane, road/car).

- The local distribution of topics that do not actually appear in an image remains uncontrolled. In the absence of information, a topic has to be located in some area of the image, with independence of the fact that it is present or not.

- RBLTM does not include useful information from global classifiers, like SVMs working on the bag-of-words, that may help to provide better segmentations.

RBLDA has been proposed in [González-Díaz and de María, 2011]. This section provides a complete description of this generative model, describing each of its constituting elements. Since the preprocessing step of RBLDA is the same as for the RBLTM the reader is referred to section 4.3.1 for a complete description.

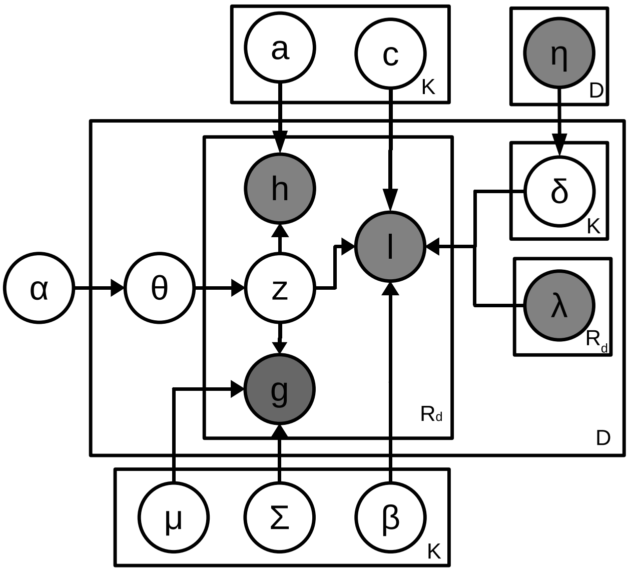


Figure 4.10: Graphical model of the Region-Based LDA. Following the standard graphical model formalism, nodes represent random variables, while the edges show the dependencies among variables. Shaded circles denote observed variables and unshaded ones denote latent variables to be inferred. Boxes refer to different instances of the same variable.

4.4.1 Description of the Generative Model of RBLDA

In this subsection we describe the structure of the generative model used in RBLDA. This model is represented in Fig. 4.10. The main contribution of this model is twofold: first, RBLDA uses an enhanced version of the location model of the RBLTM, which now considers two elements: a topic-dependent location and a context model that incorporates the influences among regions; and second, a new appearance model that successfully handles inter-word relations inside a region.

It is worth mentioning that, since this model relies on a previous blind segmentation of the image and the appearance model is associated with an entire region, the information unit is the region rather than the local patch. This constitutes a key difference between this approach and RBLTM.

We next summarize the main contributions of the proposed model with respect to LDA, on top of which is built, and RBLTM:

- In contrast to original LDA and RBLTM, there is not any variable associated with visual words, but a new appearance variable $\mathbf{h}$ that is region dependent. This new variable represents a normalized histogram of visual words within a region so that it does not depend on the number of visual words. For each region, the histogram vector has fixed length (M, the size of the vocabulary) and becomes the input of a novel distribution for the appearance.
- The variable l refers to the spatial location of a topic in an image. It basically points to the region in which the topic is located and involves two independent terms: a *topic-based term* and a *spatial context-based term.*
    - The topic-based spatial location term is computed by simply dividing an image into a fixed grid of cells and storing the probability of a topic occurring at each cell. As an example, this variable causes a topic representing 'sky' to occur more likely at the top of an image. This distribution was not included in RBLTM.
    - The context-based spatial location is image-specific and takes into account the relation between a region and its neighborhood. It provides more coherent representations so that semantically related topics tend to appear together. This variable will be later explained in section 4.4.2 and extends the prior distribution proposed for RBLTM. As an example, this variable enforces the topic 'airplanes' to occur surrounded by a semantically related topic such as 'sky'.
- In supervised environments, the model also incorporates an image-specific variable $\mathbf{g}$ that stores the outputs of other image-level classifiers (such as image-level bag-of-words classifiers).

Consequently, the generative process of the proposed model works as follows (let us note that, since the notation of LDA differs from that of PLSA, the same happens with RBLDA with respect to RBLTM):

1. For each document d, sample a Dirichlet random variable $\boldsymbol{\theta}|\boldsymbol{\alpha} \sim Dir(\boldsymbol{\alpha})$ that provides a probability distribution over the K latent topics.

2. For each region index r:

    (a) Sample a topic $z_r|\boldsymbol{\theta} \sim Mult(\boldsymbol{\theta})$.

    (b) Draw its appearance as $h_r|z_r, \mathbf{a}$ using a novel distribution to be defined in section 4.4.3.

    (c) Sample its location $l_r|z_r, \boldsymbol{\beta}, \mathbf{c}, \boldsymbol{\delta}, \boldsymbol{\lambda}$ by computing:

        i. A topic dependent spatial location $l_r|z_r, \boldsymbol{\beta} \sim Mult(\beta_{z_r})$.

        ii. A context-based spatial location (from the blind image segmentation) based on context information $l_r|z_r, \mathbf{c}, \boldsymbol{\delta}, \boldsymbol{\lambda}$, following a cooperative context distribution (described in Section 4.4.2).

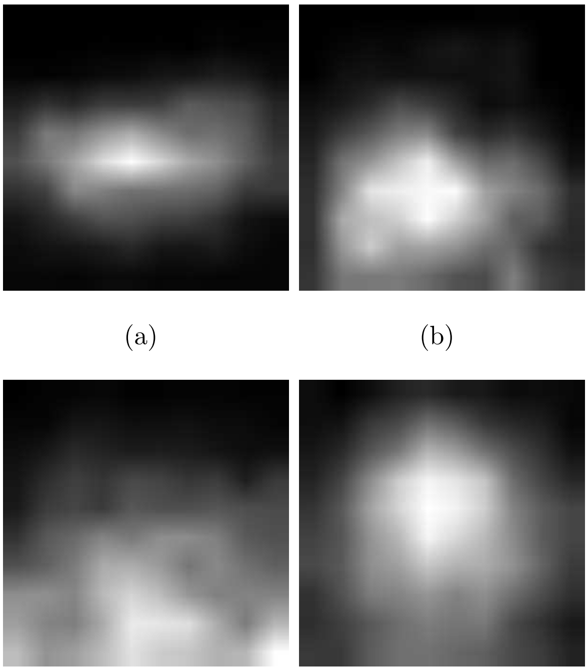
    (d) Just in a supervised environment: Sample a global variable g that depends on the outputs of a global classifier as $g|\mu, \Sigma \sim \mathcal{N}(\mu, \Sigma)$.

For the sake of compactness, we omit the subindex d in those variables that are document-dependent unless a sum over the documents is performed. Considering a particular document d in the corpus, several parameters are involved in its generative process:

- $\boldsymbol{\alpha}$ is a K-dimensional vector that contains the parameters α_k of the Dirichlet distribution (with $\alpha_k > 0, \forall k$). This parameter $\boldsymbol{\alpha}$ is shared by all the documents in the corpus.

- $\boldsymbol{\beta}$ is a collection of K N_g-dimensional vectors containing the probabilities of the topic k occurring at the N_g different cells of the grid. Since a blind segmentation method does not produce regular regions, a matching between the segmentation blobs and the cells in the image is performed by computing the proportion of the region that lies in each cell.

The rest of the model parameters will be later explained in the following subsections.



(a) (b)

(c) (d)

Figure 4.11: Some empirical distributions of the location model in RBLDA for various semantic concepts. a) Aeroplane b) Bicycle c) Dining Table d) Person

4.4.2 The location model

The location model of the RBLDA aims to estimate the spatial location of topics in an image. To achieve this purpose, it relies on two conditionally independent distributions:

1. The topic-dependent location $l_r|z_r, \boldsymbol{\beta} \sim Mult(\beta_{z_r})$, which estimates the usual location of topics in a corpus. Hence, this distribution is common for all documents in the corpus and, partitioning each image into a fixed grid of cells, it stores the probability β_{kg} of a topic k occurring at a particular cell g. Hence,

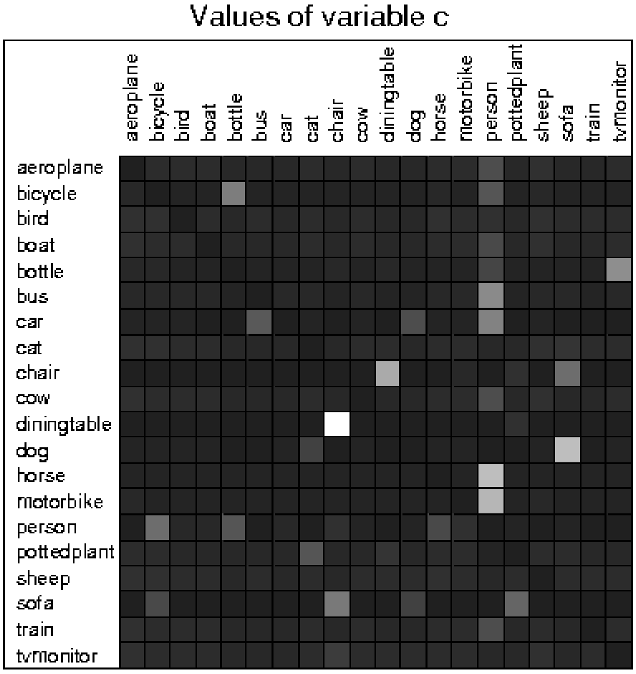


Figure 4.12: Values of the variable **c** in the context model of RBLDA for the PASCAL VOC 2010 database [Everingham et al., 2010]. **c** estimates the spatial co-ocurrence of topics in a corpus. Elements in the diagonal have been set to zero to improve the visualization.

$\boldsymbol{\beta}$ stands for a collection of K N_g-dimensional multinomial variables where N_g is the number of cells of the grid. Figure 4.11 shows some empirical examples of this distribution for various semantic concepts.

2. The context-based spatial location works on a previous segmentation of an image and studies the relationships between a region and its neighborhood. The basic idea is to locate topics in regions that are in agreement with their neighborhood (context). We consider two regions to be in agreement not only if they belong to same topic but also if they belong to correlated topics (e.g. aeroplane-sky, car-road, etc.).

The context model incorporates these relations to the generative process while keeping it simple enough to allow for closed expressions in the inference process. As mentioned before, the objective of this model is to set the basis for *inter-region inter-topic cooperation*. This means that regions belonging to a particular topic A may push other regions towards belonging to other topic B when both topics tend to appear together in the corpus. To this purpose, three variables are defined:

- $\boldsymbol{\lambda}$ stores the influences λ_{pr} between any two regions p and r in an image. Following the approach described in section 4.3.1, λ_{pr} is computed as the ratio between the common boundary of the two regions l_{pr} and the length of the region r and further normalized to obey $\sum_{r=1}^{R_d} \lambda_{pr} = 1$. This vector is precomputed and remains fixed.

- $\boldsymbol{\delta}$ is a document-dependent collection of K R_d-dimensional unknown parameters $\boldsymbol{\delta}_k = [\delta_{k1} \dots \delta_{kR_d}]$, with $\sum_{p=1}^{R_d} \delta_{kp} = 1$. Each element of the vector is called the *importance* δ_{kp} of a region p given a topic k, and must be inferred during the inference process.

- $\mathbf{c}$ stands for a collection of K K-dimensional multinomial parameters c_t such as $\sum_{t=1}^{K} c_{tk} = 1$. The objective of this variable, shared across all the documents in the corpus, is to capture the spatial correlation among topics. In other words, c_{tk} estimates the probability of co-occurrence of topics t and k in spatially adjoining regions in the corpus. Figure 4.12 shows an example for topics associated with the twenty categories in the PASCAL VOC 2010 database [Everingham et al., 2010].

Hence, the context model in our system is:

$$p(l_r|z_r, \boldsymbol{\delta}, \mathbf{c}, \boldsymbol{\lambda}) = \sum_{t=1}^{K} \sum_{p \neq l_r} c_{tz_r} \delta_{tp} \lambda_{pl_r} \tag{4.31}$$

It is noteworthy how, given a spatial location l_r, the context model considers influences from every region p in the image except for l_r. This approach avoids that the distribution c_k concentrates around the element k, what, in fact, would technically lead to an intra-topic cooperation model (the one of RBLTM). Since we still need to ensure that $\sum_s p(l_r|z_r, \boldsymbol{\delta}, \mathbf{c}, \boldsymbol{\lambda}) = 1$, this context model can be better explained as follows: given a topic k, the generative model looks for the best spatial location to draw the topic depending on the context of the regions (their neighborhood).

Adding a non-image region

The modeling of the context of a topic fails when it is not present in the image; in this case, in order that $\sum_{r=1}^{R_d} p(l_r = r|z_r, \boldsymbol{\delta}, \boldsymbol{\lambda}) = 1$, the contribution of the context term of a topic that is not in the image is unpredictable. This fact might lead to situations where topics that do not appear in the image are as much or even more likely than others that actually appear (specially, if the latter are uniformly distributed across the whole scene). To overcome this weakness, for each image d in the corpus, a new region is added, so the effective number of regions becomes $R_d^{eff} = R_d + 1$. The new region, called *non-image region* r^*, is considered to point outside the image. Consequently, it neither contains any local patch nor produces any influence on the remaining regions.

In order to locate potentially problematic topics in that non-image region we introduce some prior Dirichlet parameters $\boldsymbol{\eta}$ over the region importances $\boldsymbol{\delta}$, so that, in the absence of other information (appearance information), the topics tend to locate in the non-image region. In particular, the set-up of RBLDA with a non-image region requires to follow these steps:

1. For each image d in the corpus, the non-image region is added, so that the effective number of regions becomes $R_d^{eff} = R_d + 1$.

2. Proper values for the elements of the R_d^{eff}-dimensional $\boldsymbol{\eta}$ parameter should be provided: although the $\boldsymbol{\eta}$ parameter is document-dependent in the sense that each document has a particular number of regions R_d, this process is similar for all the images; in practice, η_r is low for $r \neq r^*$ and higher for r^*.

3. The model is trained using the EM algorithm.

In supervised environments, in which labels are provided at region level, the variable $\boldsymbol{\eta}$ is also used as region label. Once an object representing the topic k appears in an image, η_{kr} stores the proportion of the whole object that lies in the region r. In other case, if a topic is not present, the whole object lies in the non-image region.

The main consequence of this extended context model with non-image region is the inclusion of a new variational parameter $\boldsymbol{\chi}$, as shown in Fig. 4.13.

4.4.3 Improving the appearance model using a Kernel Logistic Regressor

In traditional topic models, the appearance model follows a multinomial distribution over each visual word. Although assigning topics at visual word level might seem appealing due to its simplicity, many authors work at the region level in order to provide additional descriptors that turn out to be more stable than the individual visual words (see [Shotton et al., 2009] and [Zhang and Zhang, 2004] for example). In the LDA formulation, this region-based granularity level has been traditionally handled by considering the appearance of a region as the product of the probabilities (multiplicative model) of the visual words that lie within that region. The interested reader is referred to [Cao and Fei-Fei, 2007] and [Wang et al., 2009] for more information. However, the multiplicative model may become too sensitive to a particular visual word when estimating the global probability of a region. Furthermore, this appearance model considers local patches as individual entities so that, given the topic of the region, their appearances are conditionally independent.

Our proposal is different from these approaches in the sense that a descriptor for the whole region is computed and used in the appearance model. Furthermore the appearance of a region is now modeled as a Kernel Logistic Regressor (KLR), so that this appearance model takes into account the nonlinear relations among visual words within a region. Hence, before describing in detail our proposal, we first introduce the Kernel Logistic Regressor.

Kernel Logistic Regression

The Kernel Logistic Regression is a well studied problem [Wahba et al., 1993] [Green and Yandell, 1985] [Hastie and Tibshirani, 1987]. Following the notation of the logistic regression, we consider a set of N data indexed by i so that a binary variable $y_i \in \{0, 1\}$ represents the label of the data. We aim to minimize the Negative Log Likelihood (NLL):

$$H = -\sum_{i=1}^{N}[y_i f(x_i) - \log(1 + \exp(f(x_i)))] + \frac{\lambda}{2}\|f\|^2_{\mathcal{H}_k} \tag{4.32}$$

where λ is a constant that weights the influence of the L2 regularization term and $\mathcal{H}_\mathcal{K}$ stands for the Reproducing Kernel Hilbert Space (RKHS) generated by the kernel K. Using the representer theorem, the optimal f has the form:

$$f(x) = b + \sum_{i=1}^{N} a_i K(x, x_i) \tag{4.33}$$

where a_i are the model parameters associated to each data in the dataset and b is a bias term. Some of the a_i can be zero and, even more, many of them can be set to zero without much loss of performance. Based on this fact, we can reduce the complexity of the KLR by selecting only those data that have strong influence over the final result:

$$f'(x) = b + \sum_{x_i \in S} a_i K(x, x_i) \tag{4.34}$$

where S is a subset of the training data $\{x_1, x_2, ..., x_N\}$.

The appearance model

As shown in the previous subsection, this approach is equivalent to the dual form of nonlinear support vector machines (SVMs) and takes advantage of the great discriminative power of SVMs. As shown in [Zhu and Hastie, 2001], the negative log-likelihood of eq. (4.32) has a similar shape to that of the SVM except for the well-classified samples (that still influence the KLR but no the SVM). Furthermore a KLR also provides a natural estimate of the discriminating probability $p(z_r|h_r)$, being h_r the region descriptor.

However, we use the KLR in a generative model, what induces a novel and different approach. In particular, for a given region r and topic z_r, we have proposed the use of the following distribution:

$$p(h_r|z_r, \mathbf{a}) = \frac{n_{z_r}}{1 + e^{-f_{z_r}(h_r)}} \tag{4.35}$$

with:

$$f_{z_r}(h_r) = \sum_{s=1}^{S} a_{z_r s} K(r, s) \tag{4.36}$$

where h_r represents the normalized histogram of the region r; n_{z_r} is a normalization term that ensures that $p(h_r|z_r, \mathbf{a})$ is a probability density function over the potential values of h_r; the index s points to a support point in the whole set S; $K(r, s)$ stands for the Kernel function between a region r and a support point s; and the elements a_{z_r} represent the weights of the KLR associated with the different support points. For simplicity, the bias term has been omitted. It is easy to note that, in eq. (4.36), S does not depend on z_r so that the same set of support points is used for every different KLR (every topic in the model). Hereafter, we will indistinctly use $f_{z_r}(h_r)$ and f_{rz_r}.

The normalization factor n_{z_r} in eq. (4.35) deserves some additional words. Since the combination of different words in a region leads to an infinite number of potential values for h_r, getting a proper normalization becomes unfeasible. Therefore, we have made one assumption to make this problem tractable: if our training database is large enough and contains N distinct samples, a valid normalization is achieved by simply ensuring that $\sum_{n=1}^{N} p(h_n|z, \mathbf{a}) = 1$ for each topic z. Hence, this normalization considers that no other possible combination of words can occur in the corpus, what requires that, in test, we have to assign each sample to its nearest neighbor in the training set so that unseen samples do not break the normalization.

Taking into account the negative samples

Because of the graphical model, eq. (4.35) will be evaluated for a particular region r only when the specific topic has been chosen as the one that generates the region (since z_r is an indicator variable). In a supervised framework, this issue becomes critical since the regressor in the appearance model would be trained using just those samples that are positives for the topic (being its output unknown for the negative samples). This fact would lead to situations in which a training sample that belongs to a specific topic might produce higher probabilities for other topics due to the appearance distribution.

To overcome this issue we propose the following appearance distribution to be used when training the models:

$$p(h_r|z_r, \mathbf{a}) = n_{z_r} \left(\frac{1}{1 + e^{-f_{rz_r}}} \right)^{z_r} \left(\frac{1}{1 + e^{f_{rz_r}}} \right)^{\bar{z}_r} \tag{4.37}$$

where $\bar{z}_r$ represents the complementary variable of z_r, such that $p(\bar{z}_r) = 1 - p(z_r)$.

Handling unbalanced datasets

In many cases the datasets are strongly unbalanced, i.e., the number of negatives is much higher than the number of positives for a given topic. Although the normalization term n_{z_r} may help to handle this issue, it may be useful to weight the influence of positive and negative samples separately. Of course, this idea is employed only in supervised frameworks in which the number of positive and negative samples is known a priori. This approach has its equivalent in SVMs when assigning different costs to positive and negative errors.

Our proposal is as follows: for each topic z, we compute the proportion of positive samples N_{pos} and generate the weight for the negative ones w_{z_r} as:

$$w^n_{z_r} = w_{z_r} = \frac{N_{pos}}{N} \tag{4.38}$$

Furthermore, in order to ensure that the global weight is equal for all the topics, we define the positive weight as $w^p_{z_r} = 1 - w_{z_r}$. Then, we use the following appearance distribution (just in training):

$$p(h_r|z_r, \mathbf{a}) = n_{z_r} \left(\frac{1}{1+e^{-f_r z_r}}\right)^{w^p_{z_r} z_r} \left(\frac{1}{1+e^{f_r z_r}}\right)^{w_{z_r} \bar{z}_r} \tag{4.39}$$

Obviously, in test the appearance models remain unchanged and eq. (4.35) is used.

Set of reference points

The selection of those samples that will be taken as support points in each KLR plays an important role in terms of both quality and efficiency. The first option is to use the whole training dataset, so that every region in every document is taken as reference. However, this simple approach is not computationally feasible and may lead to severe over-fitting. Hence, the main objective is to achieve an sparse representation that requires less computations and minimizes the overfitting. Various authors have investigated how to build up the set of reference points for the KLR. Several approaches use data statistics but not the label information to select the reference points ([Smola and Schökopf, 2000], [Williams and Seeger, 2001]). Others, such as [Zhu and Hastie, 2001] and [Lafferty et al., 2004], incorporate label information through greedy strategies and measure the gain coming from including a reference point.

In our proposal, we select an initial set S_0 of reference points and, at each iteration, add a new set of support points whose appearance is not well modeled yet. Additionally, in order to minimize the memory requirements, the same set of support points is shared across the whole set of topics. In particular, given the training data and a starting support set S_0, the process to select the initial set of support points is as follows (let as consider $|S|$ as the size of a set S):

1. For each topic k, consider a particular set S_k of size $|S_k| = \lfloor |S|/K \rfloor$, such that $K \cdot |S_k| \leq |S|$, where $\lfloor \cdot \rfloor$ represents an integer floor operator.

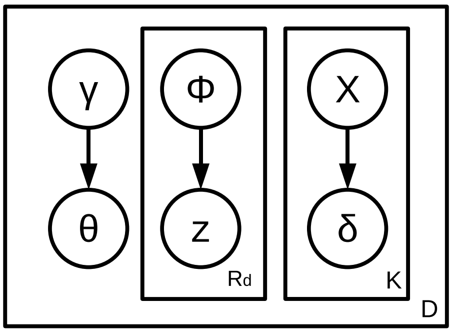


Figure 4.13: Graphical model of the variational distribution used to approximate the posterior in RBLDA. Following the standard graphical model formalism, nodes represent random variables, while the edges show the dependencies among variables. Shaded circles denote observed variables and unshaded ones denote latent variables to be inferred. Boxes refer to different instances of the same variable.

2. For each topic k, divide the positive samples into $|S_k|$ clusters and represent each cluster with the sample that is closest to the center.

3. Randomly select the remaining initial reference points $l_r = |S| - K \cdot |S_k|$.

4. At the $k-th$ iteration of the inference algorithm, a new set S_k^{new} is added to the actual set S_{k-1} by selecting the $|S_k^{new}|$ samples that show the lowest log-likelihood.

This approach is optimal when $|S_k^{new}| = 1$. In other case, some of the samples may correspond to similar cases and thus be highly correlated. However, since the appearance model is just a module in the whole generative framework, the value of this parameter can be selected as a trade-off between performance and computational complexity.

4.4.4 Inference

This Section describes the inference process. As in the original LDA, exact inference is not possible since the posterior becomes insoluble, due to coupling between the variables θ and $\mathbf{z}$. Therefore, we propose to use a simplified variational distribution q (that is tractable) and mean-field variational inference so that the Kullback-Leibler divergence between the variational distribution and the true posterior is minimized (see eq. (4.10)). The new variational distribution q is represented in Fig. 4.13 and obeys:

$$q(\boldsymbol{\theta}, \mathbf{z}, \boldsymbol{\delta}|\Theta_v) = q(\boldsymbol{\theta}|\boldsymbol{\gamma}) \prod_{r=1}^{R_d} q(z_r|\boldsymbol{\phi}_n) \prod_{k=1}^{K} q(\boldsymbol{\delta}_k|\boldsymbol{\chi}_k) \tag{4.40}$$

where $\Theta_v = \{\boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{\chi}\}$ are the variational parameters, $q(\boldsymbol{\theta}|\boldsymbol{\gamma})$, $q(\boldsymbol{\delta}|\boldsymbol{\chi})$ are Dirichlet distributions, and $q(\mathbf{z}|\boldsymbol{\phi})$ is a multinomial distribution.

Considering our parameter set $\Theta_p = \{\mathbf{a}, \mathbf{c}, \alpha, \delta, \lambda, \mu, \Sigma, \beta\}$, the new posterior can be then lower bounded as:

$$\begin{aligned}
&\log p(\mathbf{h}, \mathbf{v}, \mathbf{g}|\Theta_p) \geq E_q[\log p(\boldsymbol{\theta}|\boldsymbol{\alpha})] + \sum_{r=1}^{R_d} \Bigg(E_q[\log p(z_r|\boldsymbol{\theta})] \\
&+ E_q[\log p(h_r|z_r, \mathbf{a})] + E_q[\log p(l_r|z_r, \boldsymbol{\delta}, \boldsymbol{\lambda})] + E_q[\log p(l_r|z_r, \boldsymbol{\beta})] \\
&+ E_q[\log p(g_r|z_r, \mu, \Sigma)] \Bigg) + \sum_{k=1}^{K} E_q[\log p(\boldsymbol{\delta}_k|\boldsymbol{\eta}_k)] + H(q)
\end{aligned} \tag{4.41}$$

where $E_q[\cdot]$ denotes the expectation over the variational distribution q, and $H(\cdot)$ stands for the entropy of a distribution. An in-depth development of these and the upcoming updating formulas for the RBLDA is provided in Appendix C.

Obtaining a lower bound of the context term

The term of the log-likelihood that is associated to a region context requires computing a lower bound in order to be tractable. Hence, if we introduce a new variational parameter $r_{tkpr}/\sum_{t=1}^{K}\sum_{p=1}^{R_d} r_{tkpr} = 1$, we can apply the Jensen's inequality and get the lower bound:

$$E_q[\log p(l_r|z_r, \boldsymbol{\delta}, \boldsymbol{\lambda})] \geq \sum_{k=1}^{K}\sum_{t=1}^{K}\sum_{p=1}^{R_d} \phi_{rk} r_{tkpr} \left[\log \frac{c_{tk}\lambda_{pr}}{r_{tkpr}} + \Psi(\chi_{tp}) - \Psi\Big(\sum_{m=1}^{R_d} \chi_{tm}\Big) \right] \tag{4.42}$$

where we have additionally introduced the variational parameters $\boldsymbol{\chi}$. The term r_{tkpr} captures the whole (importance plus influence) normalized relation between two regions p and r, given that the regions p and r belong to the topics t and k, respectively.

Reducing the complexity of the appearance term

In order to reduce the complexity of the appearance term, the logistic function can be symmetrized following the approach in [Jaakkola and Jordan, 2000]:

$$\log f(x) = -log(1+e^{-x}) = \frac{x}{2} - \log(e^{x/2} + e^{-x/2}) \tag{4.43}$$

Hence, working on (4.39) gives:

$$E_q[\log p(h_r|z_r, \mathbf{a})] = \sum_{r=1}^{R_d} \Bigg\{ E_q[\log n_k] + \tag{4.44}$$
$$E_q\left[(w^p_{z_r} z_r - w_{z_r}\bar{z}_r)\frac{f_{rk}}{2}\right] - E_q\left[(w^p_{z_r} z_r - w_{z_r}\bar{z}_r)\log\left(g_{kr}\right)\right] \Bigg\}$$

where $g_{kr} = e^{\frac{1}{2}f_{rk}} + e^{-\frac{1}{2}f_{rk}}$. Since g_{kr} is convex over the variable f_k^2, the last term can be lower bounded using a first-order Taylor expansion. This process involves a new variational parameter ξ and leads to the following expression:

$$E_q[\log p(h_r|z_r, \mathbf{a})] \geq \sum_{r=1}^{R_d}\sum_{k=1}^{K} \Bigg\{ \phi_{nk}\log n_k + (\phi_{rk} - w_k)\frac{f_{rk}}{2} + \Bigg[\phi_{rk}(1-2w_k) + w_k\Bigg] \cdot$$
$$\cdot \Bigg[-\frac{\xi}{2} - \log(1+e^{-\xi_{rk}}) - A(\xi_{rk})\left(f_k^2(h_r) - \xi_{rk}^2\right)\Bigg]\Bigg\} \tag{4.45}$$

with $A(\xi_{rk}) = \frac{1}{4\xi_{rk}}\tanh\left(\frac{\xi_{rk}}{2}\right)$. Note that this lower bound is exact when $\xi^2 = f_k^2(h_r)$. Moreover, the regression function f is now outside the logarithm, thus providing a much simpler optimization.

To update the regression function, an L2-norm regularized function has to be maximized in the training phase, namely:

$$L_{f_k} = \sum_{r=1}^{R_d}\sum_{k=1}^{K} C^{(1)}_{rk} f_{rk} - C^{(2)}_{rk} f_k^2(h_r) - \frac{\mu}{2}\|f\|^2_{\mathcal{H}_k} \tag{4.46}$$

where $\mathcal{H}_{\mathcal{K}}$ stands for the Reproducing Kernel Hilbert Space (RKHS) generateed by the kernel K, and the parameters C^1, C^2 are:

$$C^{(1)}_{rk} = \frac{1}{2}(\phi_{rk} - w_k) \tag{4.47}$$

$$C^{(2)}_{rk} = [\phi_{rk}(1 - 2w_k) + w_k]\frac{1}{4\xi_{rk}} \tanh\left(\frac{\xi_{rk}}{2}\right) \tag{4.48}$$

Thus, in order to obtain the optimal parameters of the regressors $\mathbf{a}_k$, we can use an iterative Newton-Raphson method so that, at iteration t:

$$\mathbf{a}^{(t+1)}_k = \mathbf{a}^{(t)}_k - H^{-1}_k \nabla_k \tag{4.49}$$

The values of the gradient ∇_k and the Hessian H_k obey:

$$\nabla_k = K^T_k C^{(1)} - 2K^T_k (C^{(2)} \cdot f_k) - \frac{\mu}{2} K'_k \mathbf{a}_k \tag{4.50}$$

$$H_k = -2K^T_k \operatorname{diag}(C^{(2)}) K_k - \frac{\mu}{2} K'_k \tag{4.51}$$

where K and K' stand for the data Kernel matrix and the regularization matrix, respectively, and $(\cdot)$ represents the Hadamard product (element wise) between two matrices or vectors.

Parameter updating equations

To learn the values of the model parameters, we use a variational EM approach. The development of the complete formulation is provided in Appendix C. The updating equations that govern the variational parameters in the E-step of the proposed model are:

$$\xi_{rk} = \pm f_{rk} \tag{4.52}$$

$$r_{tkpr} \propto c_{tk}\lambda_{pr} \exp\big(\Psi(\chi_{kp})\big) \tag{4.53}$$

$$\chi_{kp} = \eta_p + \sum_{r\neq p}\sum_{t=1}^{K} \phi_{rk} r_{tkpr} \tag{4.54}$$

$$\begin{aligned}\phi_{rk} \propto \exp\Bigg\{ & \Psi(\gamma_k) + \log n_k - \frac{1}{2}\log|\Sigma| \\ & - \frac{1}{2}(g_k - \mu_k)\Sigma^{-1}(g_k - \mu_k) + \\ & + w_k\xi_{rk} + (2w_k - 1)\log(1 + \exp(-\xi_{rk})) + \\ & + \sum_{t=1}^{K}\sum_{p\neq r} r_{tkpr}\left[\log\frac{c_{tk}\lambda_{pr}}{r_{tkpr}} + \Psi(\chi_{tp}) - \Psi\left(\sum_{m=1}^{R_d}\chi_{tm}\right)\right]\Bigg\}\end{aligned} \tag{4.55}$$

$$\gamma_k = \alpha_k + \sum_{r=1}^{R_d}\phi_{rk} \tag{4.56}$$

In the M-step, the optimal values of the model parameters are computed. We omit the update equations for the $\boldsymbol{\alpha}$ parameter, since they do not change from the original LDA and can be found in [Blei et al., 2003]. Thus, the optimal values of the parameters to be computed in the M-step are:

$$c_{tk} \propto \sum_{d=1}^{D}\sum_{r=1}^{R_d}\sum_{p=1}^{R_d} \phi_{drk} r_{tkpr} \tag{4.57}$$

$$n_k^{-1} = \sum_{d=1}^{D}\sum_{r=1}^{R_d} \frac{1}{1 + e^{-f_k(h_{dr})}} \tag{4.58}$$

a_k, as in eq. (4.49).

$$\mu_k = \frac{\sum_{d=1}^{D}\sum_{r=1}^{R_d}\phi_{drk} g_{dr}}{\sum_{d=1}^{D}\sum_{r=1}^{R_d}\phi_{drk}} \tag{4.59}$$

$$\Sigma_k = \frac{\displaystyle\sum_{d=1}^{D}\sum_{r=1}^{R_d}\phi_{drk}(g_{dr} - \mu_k)(g_{dr} - \mu_k)^T}{\displaystyle\sum_{d=1}^{D}\sum_{r=1}^{R_d}\phi_{drk}} \tag{4.60}$$

The inference algorithm used to train the RBLDA is shown in Alg. 3.

Algorithm 3 Variational EM algorithm to perform the complete inference process in RBLDA

```
randomly initialize all the variables
Initialize the set of reference vectors $S_0$
repeat
  Variational E-Step
  for all documents $d = 1i \in D$ do
    for all regions $r \in R_d$ and topics $k \in K$ do
      compute $\xi_{rk}$, in eq. (4.52)
      repeat
        for all regions $r, p \in R_d$ and topics $k, t \in K$ do
          compute $r_{tkpr}$, in eq. (4.53)
          compute $\phi_{rk}$, in eq. (4.55)
          compute $\gamma_k$, in eq. (4.56)
          compute $\chi_{kp}$, in eq.(4.54)
        end for
      until convergence of variational procedure
    end for
  end for
  M-Step
  for all documents $d = 1i \in D$, regions $r, p \in R_d$ and topics $k, t \in K$ do
    compute $c_{tk}$, in eq. (4.57).
    compute $\mu_k$, in eq. (4.59).
    compute $\Sigma_k$, in eq. (4.60).
    add new reference points $S^n ew$ for the KLR as described in sec. 4.4.3.
    compute $a_k$, in eq. (4.49).
    compute $n_k$, in eq. (4.58).
    compute $\alpha$, as described in [Blei et al., 2003].
  end for
until convergence
```

Chapter 5

Experiments on generative models for image representation

5.1 Experimental Setup: tasks, databases, algorithms and performance measures

The performance of the proposed algorithms has been assessed in three different tasks: (i) object class segmentation: pixel-wise segmentations are generated that associate regions with object classes; (ii) image classification: a set of images is used to train the model according to a specific taxonomy and an unseen test set is then classified using the trained model; and (iii) topic discovery: semantically meaningful topics are unsupervisely discovered in a set of images.

The classification and segmentation tests have been made using the *PASCAL VOC 2010 database* [Everingham et al., 2010]. It contains 19,740 images and has been split into 50% for training/validation and 50% for testing. For classification purposes, every image in the database has an annotation file that provides a bounding box and an object class label for each of the objects. Additionally, a subset of images has been annotated pixel-wise in order that the segmentation experiments could be supported. In order to carry out the validation of several model parameters, the training set has been also divided into a train and validation sets. Hence, most of the experiments were made using the train and validation sets, whereas the final results were given on the test set, as required to establish a meaningful comparison with the official PASCAL VOC 2010 submissions. Twenty object-oriented classes were considered in the experiments (see Figure 5.4(a) for a complete list). This means that several objects from multiple classes may appear in the same image; therefore, an image may be classified as belonging to more than one category.

For the unsupervised topic discovery task evaluation, the same dataset as in the segmentation problem was used but, in this case, labels were not used so that the inference process was completely unsupervised.

In order to provide a meaningful evaluation of the proposed generative methods, we have compared their performance against several generative and discriminative approaches; in particular:

1. Dense Spatial Pyramid of Bag-of-Words model (D-BoW): It generates a dense representation of an image by computing local SIFT and color descriptors over the same dense grid as in the generative models. Then, normalized histograms are computed for both features at several spatial granularities in order to generate a spatial pyramid of histograms that is finally classified by a SVM with histogram intersection [Chang and Lin, 2001]. In particular, vertical (1,3), horizontal (3,1) and square (3,3) spatial grids have been included in our Spatial Pyramid, so that the histograms for each cell in the grid are concatenated to end up with the final input vector for the SVM. This approach has been considered only in the classification task.

2. PLSA: the fundamental algorithm on top of which RBLTM was built. Since PLSA does not model the distribution of topics along images, in the supervised environment it only uses image-based labels.

3. Spatial-LTM [Cao and Fei-Fei, 2007]: SP-LTM is a good example of supervised

region-based Latent Topic Model. It associates topics with regions rather than local patches (a region is then represented by the local patches falling into it) and uses image-level labels as priors of the document-specific topic distribution (no region-labels accepted). Our implementation differs from the original one since mean field variational methods are used for inference, rather than variational message passing, which was the learning paradigm used by the authors [Cao and Fei-Fei, 2007]

4. LDA+MRF [Zhao et al., 2010]: An extension of SP-LTM in which the authors propose the use of a Markov Random Field (MRF) to enforce spatial coherency among regions. This approach provides inter-region intra-topic cooperations.

5. Multinomial RBLDA (Mult): In order to evaluate the contribution of the advanced appearance model, this approach is similar to RBLDA but the KLR-based appearance model has been substituted by a multinomial distribution.

6. Multiclass-SVM: This approach, only used in the segmentation task, uses a multiclass SVM working on region-level histograms of words. Therefore it employs the same inputs as RBLDA and implements the multiclass classifier by means of several 1-vs-1 binary SVMs and a voting strategy [Chang and Lin, 2001].

For simplicity, LDA has been omitted in our experiments since, as stated in [Sivic et al., 2005] and confirmed in our experiments, it achieves similar results as those of PLSA in the image classification task. It is also worth noting that two of the generative models, PLSA and RBLTM, produce topics at local patch level, whereas the rest locate topics at regions that contain local patches. This difference is important for two reasons: on the one hand, locating topics at patch level is more flexible since two local patches belonging to the same region might be drawn by different topics, so the algorithms would be able to overcome deficient previous segmentations. On the other, working at region level is faster and should be more robust if proper descriptors are obtained since the model can take into account the relations between descriptors inside a region.

Table 5.1: Optimal Number of BG Topics (NoBT) for each generative model included in the classification experiments.

Generative Model	Optimal NoBT
PLSA	25
SP-LTM	5
LDA+MRF	5
Mult	5
RBLTM	20
RBLDA	4

The segmentation accuracy for a given class was assessed using the intersection/union metric, defined as the number of correctly labeled pixels of that class divided by the number of pixels labeled with that class in either the ground truth labeling or the inferred labeling.

On the other hand, the classification and topic discovery performance has been evaluated using the Average Precision (AP), a measure that has been extensively used to evaluate information retrieval systems. The AP requires a set of ranked images as system output and combines both recall- and precision-related factors in a single measure (between 0 and 1), which is also sensitive to the complete ranking. For a detailed definition of the AP measure the reader is referred to [Everingham et al., 2009].

5.2 Validating the number of BG topics

Before assessing the proposed algorithms we have previously selected a parameter of the models using the validation set: the number of topics assigned to the BG. It is noteworthy that this parameter affects to every evaluated Latent Topic Model.

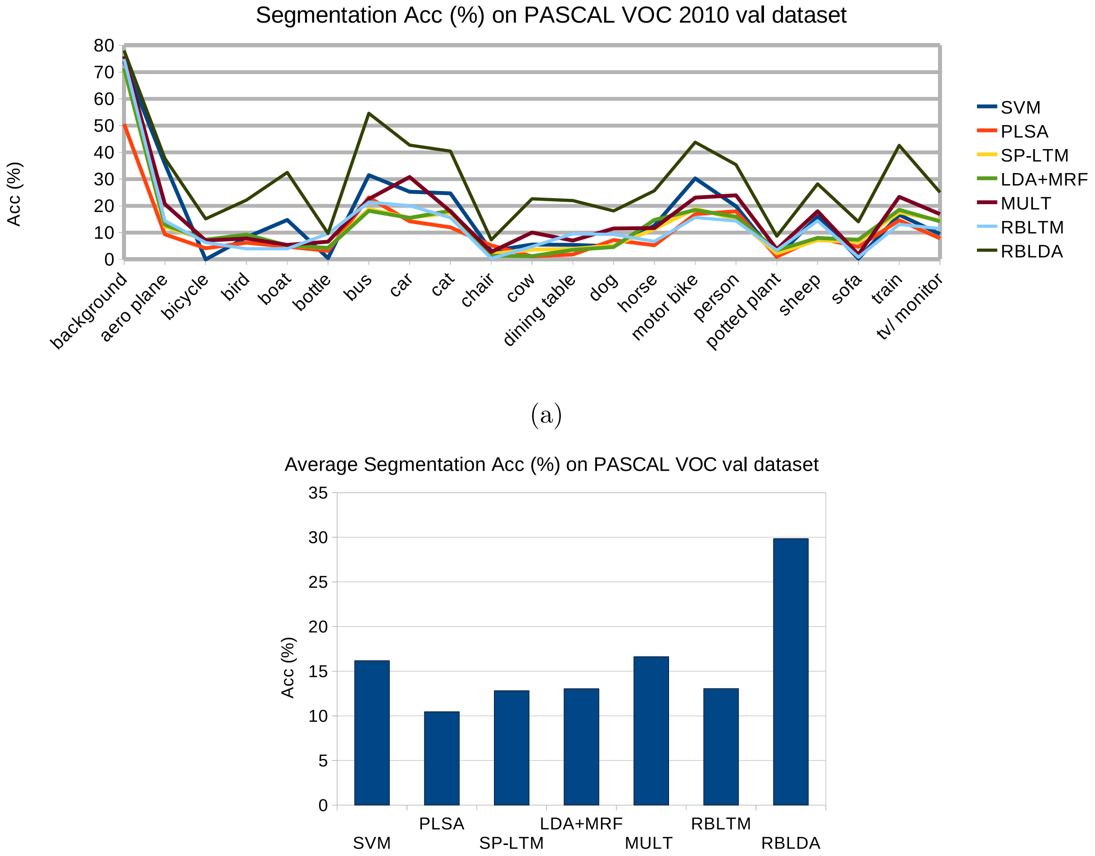
The number of BG topics deserves a few words since it plays a significant role in the representation of complex scenes with heterogeneous and normally cluttered backgrounds. On the one hand, few BG topics may lead to models that are not expressive enough to properly represent the background regions, some of which would be associated with FG topics instead. On the other hand, many BG topics may produce an unnecessary overhead in the model and, in the worst case, overfitting. Hence, in the absence of additional information, the number of BG topics was chosen by means of a cross-validation process.

Regarding the training of BG topics, it should also be noted that region-based annotations were available just for FG topics; thus, BG topics were learned from the spatial regions not labeled as FG. As a result, all the BG topics have the same prior distributions over the importances. According to this fact, it could happen that all the BG topics would be related to the same appearance vectors; however, this actually does not occur in practice if proper values for the region hyperparameters (ϵ^{i}_{REG} in RBLTM and η_{rk} in RBLDA) are selected. In particular, these parameters were chosen to be high for those FG topics that actually appear in an image, thus giving more weight to the known prior component during inference; whereas they were chosen to be low for the rest of the FG topics as well as the BG topics, thus allowing the models to behave similarly as in the unsupervised case. Hence, this configuration provides hard labels when pixel-wise segmentations of topics are available and soft labels in the absence of information.

Optimal values for the number of BG topics for each generative model were obtained and are provided in Table 5.1. It is worth noting that these results correspond to the classification task. In segmentation, although the values are slightly higher for every algorithm, similar conclusions can be drawn.

PLSA and RBLTM need many more BG topics than the rest of the models for optimal performance. We find the rationale for these results in the level with which the topics are associated. It seems that more BG topics are required when topics are assigned to regions than when they are assigned to local patches. Furthermore, the context model that stores inter-topic influences in RBLDA performs better when the number of topics is not very high so that the collection of variables $\mathbf{c}$ is not too sparse. That is the reason why RBLDA obtains the lowest value.

Additionally, we can find two specific reasons for this fact when comparing SP-LTM, Mult, and LDA+MRF with RBLDA. For the first three models, appearance probabilities of visual words are factorized in order to compute the region-level appearance probability. Working with more topics leads to more sparse multinomial appearance distributions (topics are more specific), so the factorization becomes more unstable. In comparison to RBLDA, it is worth mentioning that multinomial appearance distributions are much less expressive than our KLR-based approach, so the algorithms need more BG topics to properly model the BG of the images; whereas in RBLDA, more expressive topics are simply achieved by adding new reference vectors to the KLR, thus keeping a lower number of BG topics.



(a)

(b)

Figure 5.1: Segmentation results in terms of Average Precision (AP) achieved by all the compared algorithms for the 20 categories considered in PASCAL VOC 2010 val dataset. (a) Detailed per-category results (b) Average results.

5.3 Image Segmentation

The objective of the image segmentation task is to provide pixel-wise segmentations, i.e., an index indicative of its class is assigned to each pixel. The designed algorithms are specially appropriate for this task, due to the fact that it makes use the complete image representation that is provided by the latent topic models.

Since each generative algorithm represents images in a particular way, next, we provide some details of how to generate segmentations:

- In RBLTM, one can compute the probability $P(z_k|r_p^i)$ of a topic z_k given a region r_p^i using an equation similar to (4.17), but including in this case only those visual words belonging to the region r_p^i. As a result, the topic associated with each region of the partition, and consequently the image segmentation, could be obtained as follows:

$$z_{r_p^i} = \arg\max_{z_k} P(z_k|r_p^i) \tag{5.1}$$

- PLSA does not use segmentations. Nevertheless, since the topic is assigned at local patch level, we can provide a similar final step to that one proposed for RBLTM.

- RBLDA, SP-LTM, LDA+MRF and Mult assign topics to regions so that providing segmentations is straightforward.

Since D-BoW does not produce segmentations, it has been removed from this experiment. Instead, we have used a multiclass SVM (SVM) with histogram intersection kernel that works at region level. This supervised approach uses the same input features as the RBLDA. As mentioned before, our particular implementation of the multiclass classifier bases on one-vs-one binary classifiers and a voting strategy to select the most probable category.

Figures 5.1(a) and 5.1(b) show respectively the detailed and average segmentation results in terms of segmentation accuracy across the twenty classes and the background class. From the figures, some interesting conclusions can be drawn:

- RBLTM obtains better segmentations than its baseline algorithm, PLSA. This is a nice consequence of using the influence model and the region-based labels that are not included in PLSA. In fact, looking at the background results in Fig. 5.1(a), one can notice that the accuracy for PLSA is notably lower than for the rest of the algorithms. This is a direct consequence of using just image-labels, which leads to results in which the same topic tends to be applied to whole images, rather than dividing them into different topics. To be more precise, the appearance vector $P(w_j|z_k)$ of a FG class in RBLTM, which in this case models only regions that belong to that FG object, becomes more accurate than for PLSA.

- As one could expect, in this experiment the influence of the global probabilities g provided by D-BoW is not very strong (when compared with the classification task). Therefore, algorithms that do not use global probabilities, such as SP-LTM or LDA+MRF, achieve results that are close to those ones of the Mult, which already incorporates this information. Obviously, the influence of an image-level probability, although might help to improve the results, is not enough to make region-level decisions.

- The performance of RBLDA is clearly above those of the rest of the approaches, including Mult. This notable improvement resides in the novel KLR-based appearance distribution, which does not consider independence of visual words inside a region, as PLSA, RBLTM, SP-LTM, LDA+MRF, and Mult do. This fact provides much more expressivity and, therefore, more enhanced discrimination capabilities to the model.

- The only other method that considers relations among words inside a region is the multiclass SVM, which follows a similar approach to classify regions than the appearance model in RBLDA. Even more, the SVM has other advantages: a) it considers different sets of support vectors for each binary problem, which is not feasible in RBLDA due to its computational cost; and b) it is the only approach that is strictly multi-class, as the generative models consider conditionally independent distributions given the topic. However, looking at the results, it is noticeable that these advantages do not counteract the influence of other elements in RBLDA, such as the various background topics or the location and context terms. Furthermore, SVM results are specially worse for those categories with lower accuracies (let us say ‘difficult categories’).

- Despite the results, we can state that the most important limitation of generative models for this kind of data-driven segmentation lies in the fact that they do not conceive the segmentation problem as a pure multi-class problem in the sense that all the distributions are conditionally independent given the topic. In PLSA or RBLTM, this may lead to results in which the prior distributions of categories in the corpus are not taken into account, thus causing malfunction when the corpus is not equally distributed. Although in SP-LTM and RBLDA this issue is partially alleviated by including the corpus-based prior distributions using the hyperparameters $\boldsymbol{\alpha}$, our experiments showed that fitting the number of BG topics had a strong influence on the final results and helped to manage the balance between recall and precision.

We finally include several illustrative examples of the segmentation results for RBLTM and RBLDA in Figures 5.2 and 5.3.

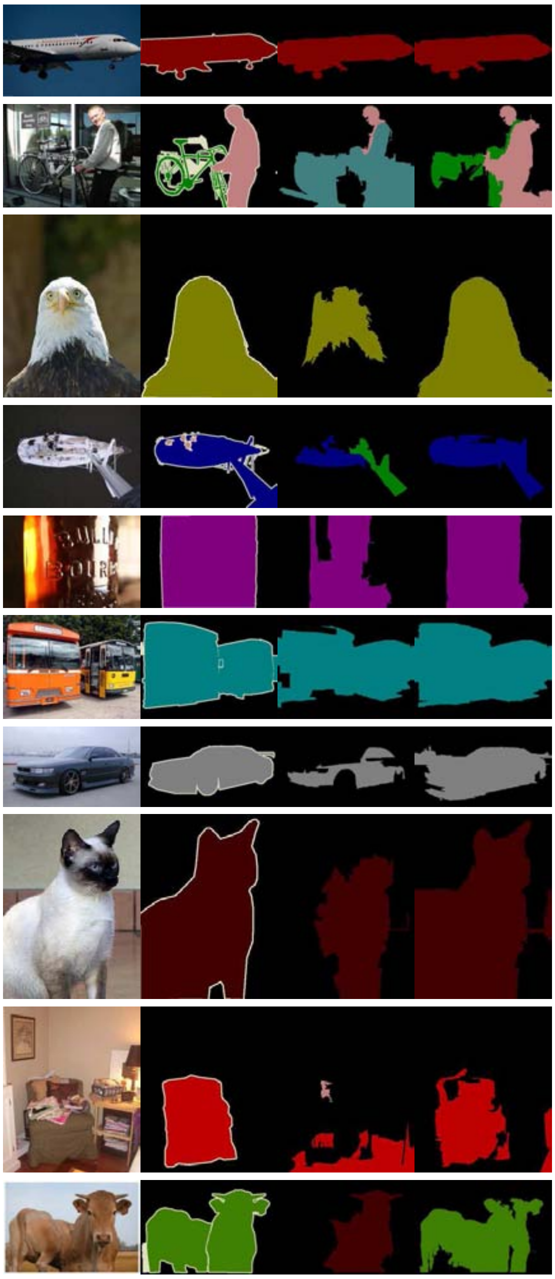


Figure 5.2: Some examples of segmentation results for categories 1-10. For each example, five images are shown. From left to right: original image, ground truth segmentation, RBLTM segmentation, and RBLDA segmentation. Each color represents a specific category; black pixels are associated with background.

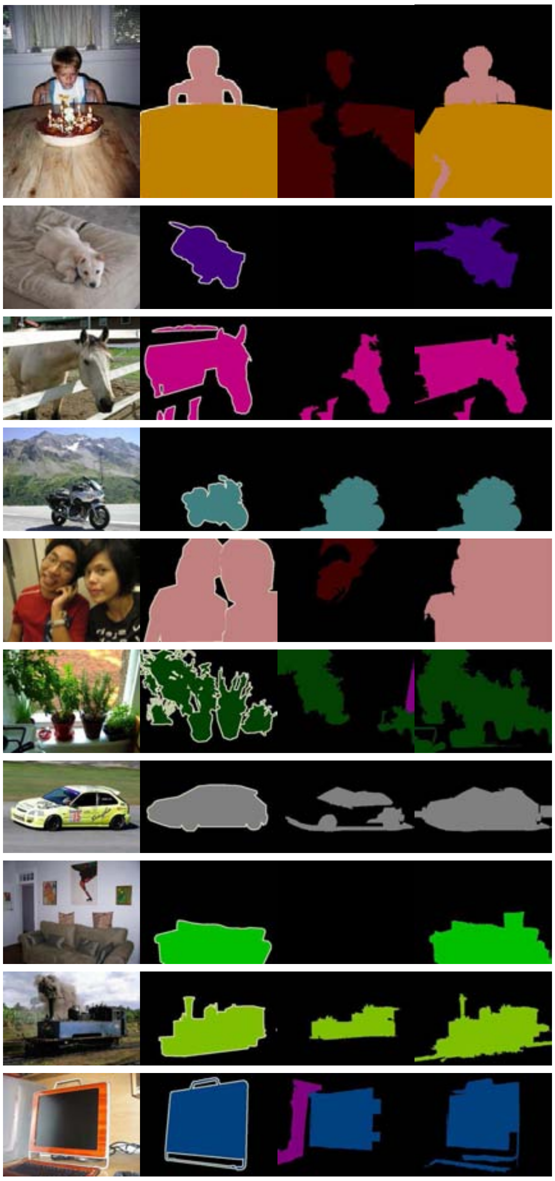


Figure 5.3: Some examples of segmentation results for categories 11-20. For each example, five images are shown. From left to right: original image, ground truth segmentation, RBLTM segmentation, and RBLDA segmentation. Each color represents a specific category; black pixels are associated with background.

5.4 Results on Image classification

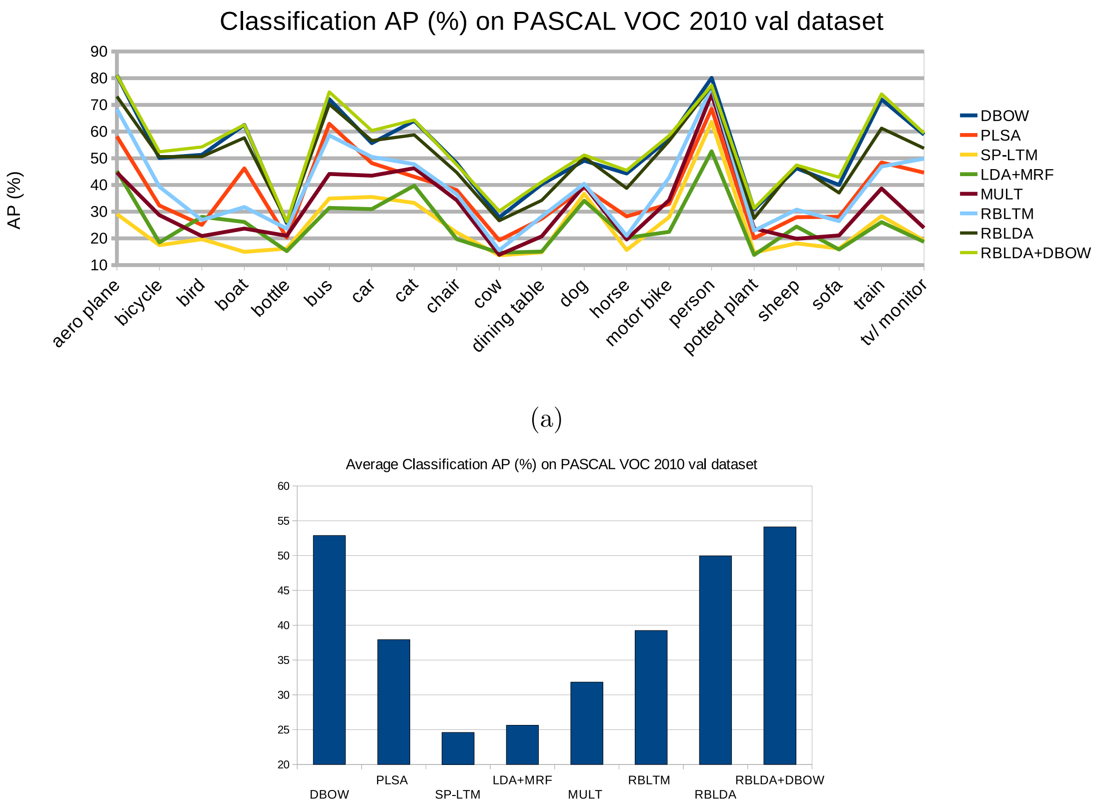
Image classification has been traditionally posed as a supervised problem in which discriminative solutions and, in particular, bag-of-words approaches, have become the prevalent technique. In such a field, generative models traditionally have not competed in terms of classification performance. However, since they provide much richer information than the presence or not of a category in an image, we have experimented with their application to this problem in order to evaluate whether they can complement discriminative models and enhance their performance.

For each image in the training set, a mask was generated from the bounding box to compute the region labels that have been used either by RBLTM and RBLDA. Then, supervised versions of the algorithms were used to optimize the model parameters, using for this purpose the corresponding foreground topic for each class (FG topic) and several background topics (BG topics).

In order to evaluate the performance of the models, the resulting generative probabilities at document level ($P(z_k|d_i)$ for RBLTM and PLSA, and γ_d for RBLDA or SP-LTM) for those topics associated with FG classes are used as estimates of the probability of a class occurring in an image. Furthermore, in order to establish a fair comparison among the algorithms, the global input vectors g have been removed from RBLDA and SP-LTM. This modification allows us to compare the D-BoW and the generative model, and helps us to evaluate the influence of the new elements in RBLDA. Furthermore, another RBLDA version (RBLDA+D-BoW) includes the global variable from D-BoW and shows the performance of the combination of both generative and discriminative approaches.

Figure 5.4 shows the AP classification results for all the algorithms included in the experiments. From the figure, interesting conclusions can be drawn:

- RBLTM slightly outperforms PLSA, the basic algorithm on top of which was built. Again, this result is supported by two reasons. First, the region-based annotation used by RBLTM provides relevant clues for image classification.



(a)

(b)

Figure 5.4: Classification results in terms of Average Precision (AP) achieved by all the compared algorithms for the 20 categories considered in PASCAL VOC 2010. (a) Detailed per-category results (b) Average results.

Second, RBLTM also incorporates an active model for region interaction that enhances its performance when dealing with regions that cannot be longer classified from their appearance; in those cases, the surrounding regions push the considered region toward spatially coherent topics.

- On the other hand, the use of region-based labels in RBLTM may also turn out to be counterproductive for image classification: since appearance vectors in PLSA integrate both FG and BG regions (PLSA does not use region-based labels), this algorithm captures the relations between highly correlated categories and backgrounds (aeroplane/sky, car/road, horse/grass, etc.), what does not happen in RBLTM. Hence, since the intra-topic inter-region cooperative model in RBLTM does not takes into account relations among topics, it becomes a clear disadvantage of RBLTM. However, overall, RBLTM still outperforms PLSA.

- SP-LTM achieves very poor results. One of the reasons was already discussed for the previous case: the lack of region annotations in the training phase. A second reason lies in the appearance distributions, which dramatically degrade the performance when a non-likely local descriptor appears in a region. Furthermore, given a region, all the visual words that lie inside are considered to provide the same influence over the final result, so that the multinomial appearance distributions do not fit the data properly.

- LDA+MRF, although improves the results of SP-LTM by means of the inclusion of a MRF that imposes some kind of spatial coherence, still suffers from the aforementioned drawbacks (absence of region annotations and multinomial model). Moreover, the better results obtained by Mult are caused by the inclusion of region annotations in the training phase, since Mult also employs the same multinomial distribution of SP-LTM and LDA+MRF.

- RBLDA achieves much better results than the rest of the generative models. The reasons are several and rely on the new elements that this model incorporates: the KLR-based appearance model is much more expressive and discriminative than the multinomial models and takes into account the relations among visual words inside a region; the context model allows for inter-topic inter-region relations, thus overcoming the mentioned issue for region-labeled approaches (lack of relation between foreground objects and their background), etc.

- None of the generative models obtains better results than the discriminative approach included for comparison. Hence, our results agree with the fact, already pointed out by other authors (see [Lazebnik et al., 2006] for an image classification example), that discriminative solutions working with whole histograms generally outperforms basic generative models in classification tasks. Thus, considering global histograms that incorporate all the information in the image normally outperforms any in-depth spatial analysis of the image. Nevertheless, the generative approaches provide much richer information concerning the underlying structure and the semantic of an image.

- The last idea is supported by the fact that the version of RBLDA that includes the outputs of the D-BoW shows the best performance in our experiments, what means that the information provided by RBLDA actually complements BoW models.

Finally, Figure 5.5 shows illustrative examples of classification results achieved by the proposed method.

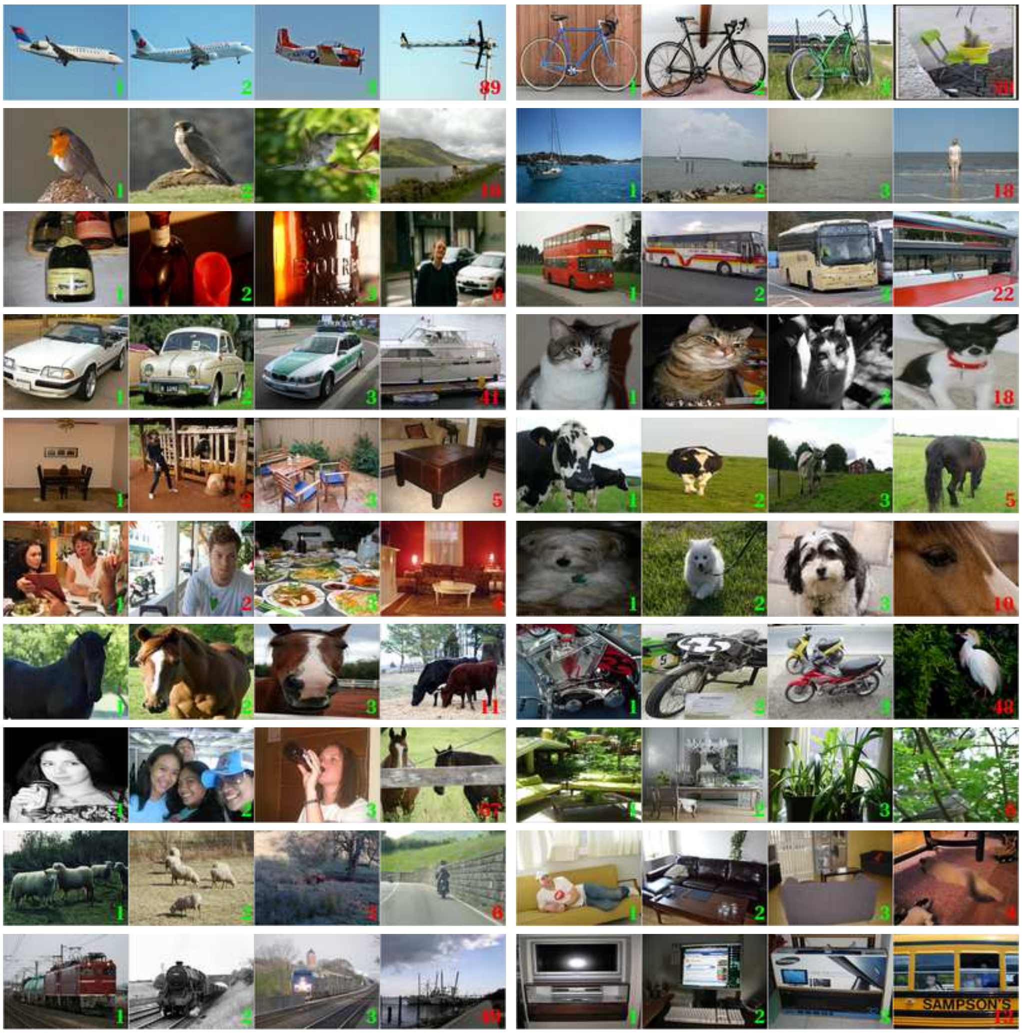


Figure 5.5: Some examples of classification results achieved by the RBLDA+D-BoW approach. Each row shows retrieved images for two categories. The number in the lower right corner indicates the place of the retrieved image in the ranking; the color of the number tells if the result is correct (green) or incorrect (red). The first three images of each row are the top 3 retrieved images, while the fourth shows the first error outside the top 3 results.

5.5 Topic Discovery

The topic discovery experiment evaluates the ability of Latent Topic Models to unsupervisely detect semantic concepts in images. To that end, unsupervised versions of the algorithms have been run over a dataset containing 964 images (the segmentation val dataset) showing twenty object-oriented categories. Since, on the one hand, each image may contain more than one category and, on the other, show many other elements that are not considered as a category, the number of latent topics K is usually higher than the number of categories C. In particular, it has been set to 50 in our experiments.

Following a similar approach as in the classification task, a vector of topic probabilities is obtained for each document in the dataset. Then, each category is assigned to the most likely topic. The way we do this alignment is as follows:

- For each pair of category c and topic k, we compute the Average Precision $AP(c,k)$. This step produces an AP matrix of size CxK.
- We select the indexes $\{c,k\}$ associated to the maximum value of the AP matrix, and assign this AP value with the selected category c.
- We set to zero all the values in the row c and the column k.
- We proceed in the same way until all the categories are assigned to one topic. The rest of the topics are therefore assigned to background elements in images.
- We compute the average AP using the values obtained for each of the FG categories.

As shown in Table 5.2, this experiment draws surprising results that do not agree with what we have seen so far. For this specific task, simple appearance models turn out to work better. Furthermore, we do not appreciate meaningful differences between RBLTM (the best option), PLSA, and SP-LTM. However, the performance of RBLDA is very poor when compared to the rest of the algorithms. The rationale

Table 5.2: Topic Discovery Results in AP for the considered Latent Topic Models

Latent Topic Model	AP (%)
PLSA	24.96
SP-LTM	24.02
LDA+MRF	21.94
Mult	23.23
RBLTM	25.15
RBLDA	19.31

behind this fact is that the KLR-based appearance model of RBLDA is probably too complex and does not fit well with usual clustering assumptions. In particular, we perceive that similar inputs are not well clustered into the same topic due to the fact that kernel measures decrease dramatically when two vectors are not very close in the feature space. This situation, although provides great precision in a (semi)supervised framework, becomes a clear limitation in an unsupervised one.

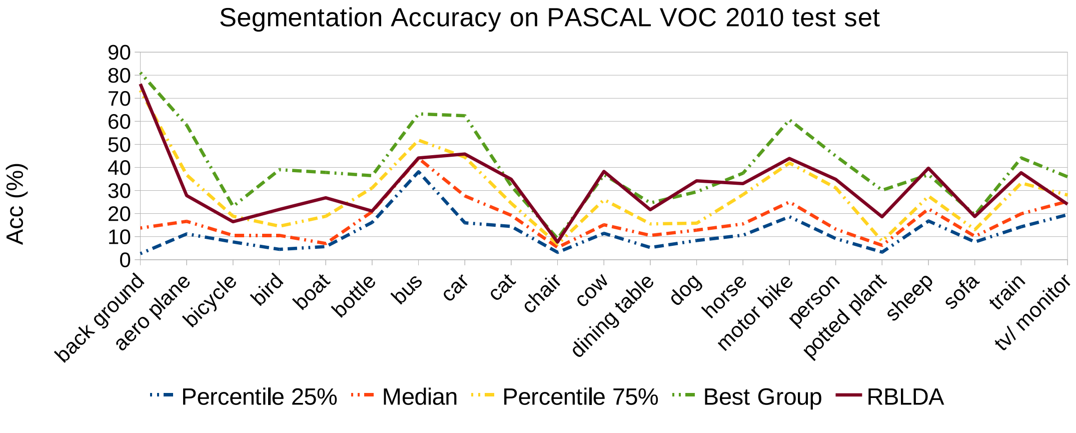
Additionally, it should be noted that RBLTM slightly outperforms PLSA, mainly due to the cooperative model that yields more realistic representations of images as mixture of topics (as already mentioned, PLSA tends to assign just one topic to each image). SP-LTM, due to the factorization of appearance probabilities, performs a bit worse than the other two. Furthermore, applying terms enforcing spatial coherence does not produce good results in unsupervised environments, as the results of LDA+MRF or Mult demonstrate.

5.6 Where we are: a comparison against official PASCAL VOC 2010 submissions

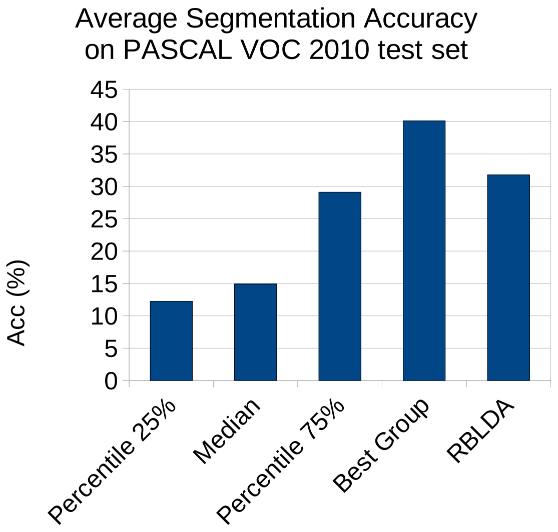
In this section we will compare our best performing proposal, RBLDA, against official PASCAL VOC 2010 submissions. As mentioned before, PASCAL VOC is probably the main challenge for several topics in Computer Vision: image classification, object detection, category-based segmentation, etc. It is worth noting that many of the proposals are not just algorithms, but complex systems with many features and classifiers. As an example, the winning group in the PASCAL VOC 2010 classification task [Chen et al., 2010] employs a large set of low-level features at several granularities: SIFT, LBP [Ojala et al., 1996], HoG [Dalal and Triggs, 2005], GIST [Oliva and Torralba, 2001], etc. This means that, in general, these systems outperform individual algorithms such as the ones presented in this thesis, that simply employ SIFT and color descriptors. Anyway, we think that establishing this comparison may help to place our developments into the state-of-the-art in Computer Vision.

Figure 5.6 shows a comparison of RBLDA and several statistics in PASCAL VOC 2010 segmentation task. Results demonstrate that our algorithm achieves a great performance in segmentation, ranking above the 75% percentile of the submissions.

Furthermore, Figure 5.7 shows a similar comparison in PASCAL VOC 2010 classification task. As shown in the figure, our proposal results are similar to the median of the submissions, what validates our approach.

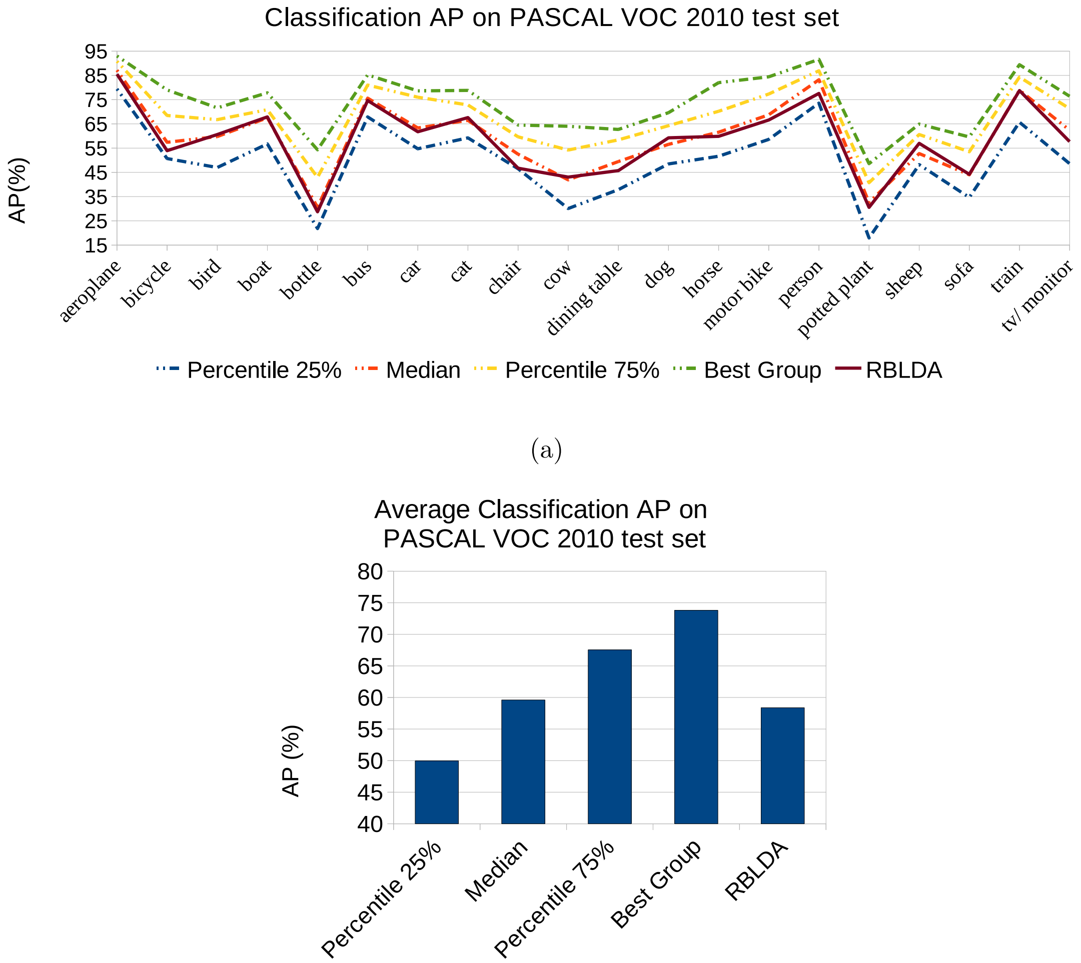


(a)



(b)

Figure 5.6: A comparison of RBLDA against PASCAL VOC 2010 official segmentation submissions. a) Detailed per-category results. b) Average results. Best group can be found in [Boix et al., 2010].



(a)

(b)

Figure 5.7: A comparison of RBLDA against PASCAL VOC 2010 official classification submissions. a) Detailed per-category results. b) Average results. Best group is [Chen et al., 2010].

5.7 Summary

In this chapter of the thesis we have assessed the performance of our two proposals in comparison to other generative and discriminative approaches. Three tasks of special interest in Computer Vision have been addressed in the evaluation: category-based image segmentation, image classification, and topic discovery.

In general, we can conclude that RBLDA works very well in supervised environments due to several extensions that are not modeled by the rest of the algorithms and, in particular, the KLR-based appearance model. Furthermore, we have demonstrated that generative models complement discriminative approaches and that the combination of both obtains the best performance.

However, in unsupervised environments, simple appearance models such as the ones in RBLTM or PLSA achieve better results and turn to be more suitable for clustering problems.

Finally, we have also assessed our developments in comparison to the official submissions of the PASCAL VOC 2010 challenge, obtaining excellent results for image segmentation and good results in image classification.

Chapter 6

Conclusions and future lines of research

6.1 Conclusions

In this thesis we have proposed a set of generative models to address two well-known but yet unsolved problems in Computer Vision: image and video segmentation and image representation.

There are two approaches to image segmentation: a) blind unsupervised techniques that generate image partitions using available information such as texture, color, or motion; and b) supervised category-based image segmentation approaches that not only generate the partitions, but also assign each pixel in the image to a particular semantic concept.

In Chapter 2, we have proposed an algorithm for blind segmentation that takes an input video sequence and produces image segmentations of the keyframes. The algorithm exploits spatio-temporal information of the video in terms of color, spatial, and motion features to generate coherent segmentations in which regions are closer to semantic objects. The algorithm relies on the Mixture of Gaussians (MoG), a well-known clustering technique, in which several prior distributions have been introduced to end up with an iterative adaptive clustering algorithm that produces new regions at each iteration until convergence. Our prior distributions are based on a diagonal matrix of hyperparameters, so that the balance of known (from previous iterations) and unknown distributions is individually managed for each feature. The individual balance is interesting since some features, like color, require less adaptation once a region has been marked to be added. In contrast, if other features like the spatial location receive a higher degree of freedom, they will enhance the adaptation of the component to the region boundaries.

Furthermore, a so-called K-management module was designed to decide whether or not to add new classes at each iteration of the algorithm. This module takes advantage of several novel spatio-temporal mid-level features that model geometric properties of real world objects and motion patterns.

In Chapter 3 the proposed algorithm has been assessed in a very challenging database built from clips of the TRECVID 2006 database. This database is not specific for segmentation but for multimedia information retrieval, a field that is more general and for which our algorithms are intended. The experiments shown that our proposal outperforms other approaches in the literature and specially emphasized the performance increase due to the spatio-temporal features, at both low and mid-level. However, the objective evaluation measures turned out to be very little discriminative when comparing visually different segmentations, thus providing not very meaningful comparisons among the involved algorithms.

Blind segmentation algorithms become one of the inputs for the generative models proposed in Chapter 4. In this chapter, after providing an in-depth review of well known techniques for image recognition, object discovery, and category-based segmentation, we have proposed two Latent Topic Models for image representation. The first one, named RBLTM, extends PLSA to model the spatial location of topics in images. This objective has been fulfilled by incorporating a cooperative distribution that not only assigns semantic concepts to specific regions in the image, but also exchanges information among regions so that topics are successfully arranged in the image. Furthermore, we have also provided a formal framework for supervised training that is able to manage both image-level labels and pixel-wise segmentations.

The second approach, known as RBLDA, has been conceived to overcome several particular drawbacks of RBLTM. First of all, RBLDA has been adapted to the LDA formulation, which turns to be a more Bayesian approach that learns corpus-level topic distributions that are not modeled in PLSA. Second, RBLDA models the nonlinear relations between visual words that lie in the same region by computing new descriptors at region level. Specifically, a novel KLR-based appearance model is able to exploit the potential of the region descriptors, thus becoming much more discriminative. Third, RBLDA provides an extension of the spatial location model in which two conditionally independent distributions arise: a topic-dependent location model that stores image locations in which a topic tends to appear, and a context model that now allows for inter-region inter-topic cooperation (in contrast with the inter-region intra-topic model of the RBLTM). Finally, RBLDA also allows us to use image-level topic probabilities built from other global classifiers (such as discriminative Bag-of-Words approaches).

In Chapter 5 we have compared the proposed algorithms to some basic Latent Topic Models approaches and other state-of-the-art alternatives. Using a very challenging up-to-date database such as the PASCAL VOC 2010 database, our proposals were tested in three different tasks: image classification, category-based image segmentation and object discovery. For both image classification and category-based segmentation, RBLDA turned to be the most powerful approach, mainly due to the advanced appearance distribution as well as the inter-region inter-topic cooperation model. RBLTM, however, although outperforms its reference approach (PLSA), does not achieve comparable results to RBLDA.

Furthermore, even though state-of-the-art discriminative approaches still overcome generative models for image classification, we have proved that how the combination of both reaches the best performance. This result allows us to draw an interesting conclusion: generative and discriminative models provide complementary information and, consequently, they can be successfully combined. In RBLDA, this combination is natural due to the aforementioned global image-level variable provided by the model which simply plugs the discriminative information into the generative approach. For image segmentation RBLDA outperforms SVM-based discriminative approaches, notwithstanding, owing to the remainder elements in the model (context, location, etc.).

On the other hand, the experiments for unsupervised object discovery led to different results. In this case, the models should detect latent topics that explain the generation of documents and where the complex models used in RBLDA do not work properly. In fact, for this task RBLTM yielded the best performance, whereas RBLDA achieved the worst results.

Hence, we can conclude that each algorithm is specifically well-suited for a particular situation, mainly defined by the available information and the degree of supervision.

Finally, we also have compared our best performing method for image classification and category-based image segmentation (RBLDA) to the official PASCAL VOC 2010 submissions, where it achieves notable results in image classification and excellent in segmentation.

6.2 Future lines

Here we sketch some of the most promising lines of research that arise from this thesis.

Regarding the spatio-temporal segmentation algorithm, the most evident direction of the research involves adapting the algorithm for object tracking in video sequences. Hence, the same adaptive probabilistic framework that supports the splitting process can be used to track the partition in subsequent frames. At the same time, novelty detection techniques should be studied to discover new objects appearing in the scene. Furthermore, the use of mid-level features can also be extended to support plausible transformations of the regions from one frame to the next.

The advantages of this novel approach for the object tracking task are diverse: refinement of the segmentation using information from several frames, estimation of the motion patterns of the objects in a scene, spatio-temporal characterization of objects for classification or video object segmentation for coding purposes.

Furthermore, it will be also nice to provide an evaluation on a multimedia information retrieval task to check whether the proposed solution can be helpful in tasks like semantic concept detection in video. This approach, although has been followed in our submission for TRECVID 2009 [González-Díaz et al., 2009b], was embedded in a complex system in which the contribution of this subsystem was not assessed.

There is also room for improving the proposed generative models for image representation. The simplest extension entails the addition of new features such as region shape features. More advanced developments may incorporate new elements such as part-based models, which have demonstrated to be very discriminative (see [Felzenszwalb et al., 2010b, Felzenszwalb et al., 2010a] for two good examples).

Looking at the theoretical framework, new levels in the hierarchy can be added to end up with very interesting generative models such as the Hierarchical Dirichlet Processes (HDP) [Teh et al., 2006], which exploit correlations among different corpus of data.

Finally, one very promising line of research is the adaptation of the latent topic models to dynamic scenarios involving video sequences. The main objective of this approach is to model the spatio-temporal properties of the semantic video concepts and therefore enhance the detection and segmentation performance.

Appendices

Appendix A

Derivation of the formulas for the adaptive probabilistic clustering

A.1 Expansion of the Maximum a Posteriori (MAP)

In this appendix, we derive the final equations of the adaptive Mixture of Gaussians (MoG) used in the spatio-temporal segmentation algorithm. Along this section, for the sake of compactness, the actual index ranges are just given here and omitted later on: $i = 1, 2, ..., N$, $j, k = 1, 2..., K,$.

As mentioned in section 2.3.5, the posterior follows the next expression:

$$p(x, \theta) = \left[\sum_k \alpha_k N(x|\mu_k, \Sigma_k) \right] p(\theta|\theta_0) \tag{A.1}$$

where α_k are the mixing coefficients of the MoG, N is the normal distribution and θ is the parameter set $\theta = \{\alpha_k, \mu_k, \Sigma_k, k = 1...K\}$. Moreover, θ_0 represents the prior knowledge about the parameters. The inclusion of priors leads the EM algorithm to find a Maximum A Posteriori (MAP) rather than the Maximum Likelihood (ML) values of the parameters. Taking logarithms and extending this equation gives:

$$\begin{aligned}\log p(x,\theta) = \quad & \textstyle\sum_i \log\left(\sum_k \alpha_k N(x|\mu_k,\Sigma_k)\right) + \log p(\alpha|\alpha_0) \\ & + \textstyle\sum_k \log p(\Sigma_k|\Sigma_{0k}, m_k) + \sum_k \log p(\mu_k|\mu_{0k},\Sigma_k, M_{\beta k})\end{aligned} \tag{A.2}$$

which can be solved using the EM algorithm.

A.2 Expectation Step

The first term in eq. (A.2) corresponds to the classical MoG paradigm, which can be solved introducing a new term r_{ik} and computing a lower bound of the form:

$$\begin{aligned}&\sum_i \log\left(\sum_k \alpha_k N(x|\mu_k,\Sigma_k)\right) \geq \sum_{i,k} r_{ik} \log\left(\frac{\alpha_k N(x|\mu_k,\Sigma_k)}{r_{ik}}\right) \\ &\propto \sum_{i,k} r_{ik}\left[\log\alpha_k - \frac{1}{2}\log(2\pi|\Sigma_k|) - \frac{1}{2}(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)\right]\end{aligned} \tag{A.3}$$

The new term r_{ik} stands for the posterior probability of a data x_i being sampled from the kth component of the mixture. Computing the derivative of this term with respect to r_{ik} and adding a Lagrange multiplier that ensures that $\sum_k r_{ik} = 1$ provides the following update equation:

$$r_{ik} = \frac{\alpha_k p(x_i|\mu_k,\Sigma_k)}{\sum_{j=1}^K \alpha_j p(x_i|\mu_j,\Sigma_j)} \tag{A.4}$$

A.3 Maximization Step

In this section, we derivate the expressions of the model parameters $\{\alpha,\mu,\Sigma\}$ making use of the prior distributions proposed in eq. (2.5), (2.7) and (2.6), respectively.

A.3.1 Mixing proportions

If we make use of the expansion provided in eq. (A.3) as well as the definition of the Dirichlet prior in eq. (2.5), and add a Lagrangian multiplier λ that ensures

$\sum_k \alpha_k = 1$, we can then write the terms of the log-posterior that contain α:

$$\log p_\alpha \geq \sum_{i,k} r_{ik} \log \alpha_k + \sum_k (\alpha_{0k} - 1) \log \alpha_k - \lambda \sum_k (\alpha_k - 1) \tag{A.5}$$

The derivative of this expression obeys:

$$\frac{\partial \log p_\alpha}{\partial \alpha_k} = \sum_i r_{ik} \frac{1}{\alpha_k} + (\alpha_{0k} - 1) \frac{1}{\alpha_k} - \lambda \tag{A.6}$$

Setting this derivative to zero and computing the value of the multiplier we gives update expression for the mixing proportions:

$$\alpha_k = \frac{\sum_i r_{ik} + (\alpha_{0k} - 1)}{n + \sum_j (\alpha_{0j} - 1)} \tag{A.7}$$

where n represents the total number of pixels in the image.

Finally, if we substitute the hyperparameters for those expressions previously proposed in eq. (2.8) we get the final equation that updates the mixing proportions:

$$\alpha_k = \frac{\sum_{i=1}^{n} r_{ik} + \frac{1}{d} Tr(M_{\beta k}^{-1} M_{\beta k}^{-1}) \alpha'_k}{\sum_{j=1}^{K} \frac{1}{d} Tr(M_{\beta j}^{-1} M_{\beta j}^{-1}) \alpha'_j + n} \tag{A.8}$$

where d stands for the dimension of the feature space.

A.3.2 Gaussian Means

As already mentioned, the mean is normal with a transformation matrix $M_{\beta k}$ (see eq. (2.7)). The terms of the log-posterior that include the Gaussian mean are shown in the following equation:

$$\begin{aligned} \log p_\mu \quad &= -\tfrac{1}{2} \textstyle\sum_{i,k} r_{ik} (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) \\ &-\tfrac{1}{2} (\mu_k - \mu_{0k})^T \left(M_{\beta k} \Sigma_k M_{\beta k}^T \right)^{-1} (\mu_k - \mu_{0k}) \end{aligned} \tag{A.9}$$

Again, computing the derivative gives:

$$\frac{\partial \log p_\mu}{\partial \mu_k} = - \sum_i r_{ik} \Sigma_k^{-1} (x_i - \mu_k) + M_{\beta k}^{-1} \Sigma_k^{-1} M_{\beta k}^{-1} (\mu_k - \mu_{0k}) \tag{A.10}$$

which gives place to the following update equation when set to zero:

$$\mu_k = M_A^{-1}\left(\sum_i r_{ik} M_{\beta k} x_i + M_{\beta k}^{-1}\mu_{0k}\right) \tag{A.11}$$

where $M_A = \left(\sum_{i=1}^n r_{ik} M_{\beta k} + M_{\beta k}^{-1}\right)$.

The final update equation comes from substituting the hyperparameter μ_{0k} as mentioned in eq. (2.8):

$$\mu_k = M_A^{-1}\left(\sum_{i=1}^n r_{ik} M_{\beta k} x_i + M_{\beta k}^{-1}\mu'_k\right) \tag{A.12}$$

As it can be noticed, the mean is conditioned on the covariance matrix, so that the latter should be computed before during the inference.

A.3.3 Gaussian Covariances

The inverse of the covariance matrix (precision)is Wishart with m_k degrees of freedom as defined in eq. (2.6). We therefore expand the terms of the log-posterior that depend on Σ_k:

$$\begin{aligned}\log p_\Sigma \quad &= -\tfrac{1}{2}\textstyle\sum_{i,k} r_{ik}(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k) - \tfrac{m_k-d-2}{2}\log|\Sigma_k| \\ &-\tfrac{1}{2}Tr\left(\Sigma_{0k}\Sigma_k^{-1}\right) - \tfrac{1}{2}(\mu_k-\mu_{0k})^T\left(M_{\beta k}\Sigma_k M_{\beta k}^T\right)^{-1}(\mu_k-\mu_{0k})\end{aligned} \tag{A.13}$$

We compute the derivative as:

$$\begin{aligned}\frac{\partial \log p_\Sigma}{\partial \Sigma_k} &= -\frac{1}{2}\sum_i r_{ik}\left[\Sigma_k^{-1} + \Sigma_k^{-1}(x_i-\mu_k)(x_i-\mu_k)^T\Sigma_k^{-1}\right] \\ -\frac{m_k-d-2}{2}\Sigma_k^{-1} &+ \frac{1}{2}\Sigma_k^{-1}\Sigma_{0k}\Sigma_k^{-1} + \frac{1}{2}\Sigma_k^{-1}M_{\beta k}^{-1}(x_i-\mu_k)(x_i-\mu_k)^T M_{\beta k}^{-1}\Sigma_k^{-1}\end{aligned} \tag{A.14}$$

Again, we set this derivative to zero and get the update expression:

$$\Sigma_k = \frac{\sum_i r_{ik}(x_i-\mu_k)(x_i-\mu_k)^T + M_{\beta k}^{-1}(x_i-\mu_k)(x_i-\mu_k)^T M_{\beta k}^{-1} + \Sigma_{0k}}{\sum_i r_{ik} + m_k - d} \tag{A.15}$$

Again, substituting the hyperparameter Σ_{0k} by the proposed value in eq. (2.8) gives:

$$\Sigma_k = \frac{\sum_{i=1}^{n} r_{ik}(x_i - \mu_k)^2 + M_{\beta k}^{-1}((\mu_k - \mu'_k)^2 + \Sigma'_k)M_{\beta k}^{-1}}{\sum_{i=1}^{n} r_{ik} + \frac{1}{d}Tr(M_{\beta k}^{-1}M_{\beta k}^{-1})} \tag{A.16}$$

Appendix B

Derivation of the formulas for the RBLTM

B.1 Derivation of the formulas for the unsupervised RBLTM

In this appendix, we derive the equations of RBLTM that differ from those of PLSA. Along this section, for the sake of compactness, the actual index ranges are just given here and omitted later on: $i = 1, 2, ..., D$, $j = 1, 2..., M$, $l, p = 1, 2, ..., R_i$, and $k = 1, 2, ..., K$.

B.1.1 Computation of a lower bound of the log-likelihood to obtain the posterior probabilities of the latent variables

We start from the likelihood function given in (4.13):

$$L = P(X|\theta) = \prod_{i=1}^{D} \prod_{j=1}^{M} \prod_{l=1}^{R_i} P(d_i, w_j, s_l^i)^{n^i(w_j, s_l^i)}, \tag{B.1}$$

take logarithm and expand terms to obtain:

$$\log L = \sum_{i,j,l} n^i(w_j, s_l^i) \left[\log P(d_i) + \log \sum_k P(z_k|d_i) P(w_j|z_k) P(s_l^i|z_k, d_i, \boldsymbol{\alpha}) \right] \tag{B.2}$$

Now, rewriting the previous equation by introducing $P(z_k|d_i, w_j, s_l^i)$ (which obeys $\sum_{k=1}^{K} P(z_k|d_i, w_j, s_l^i) = 1$) and applying the Jensen's inequality, the following lower bound of the log-likelihood is obtained:

$$\log L \geq \sum_{i,j,l} n^i(w_j, s_l^i) \left[\log P(d_i) + \sum_k P(z_k|d_i, w_j, s_l^i) \log \frac{P(z_k|d_i) P(w_j|z_k) P(s_l^i|z_k, d_i, \boldsymbol{\alpha})}{P(z_k|d_i, w_j, s_l^i)} \right] \tag{B.3}$$

Computing the derivative of the log-likelihood with respect to $P(z_k|d_i, w_j, s_l^i)$ yields the equation (4.16):

$$P(z_k|d_i, w_j, s_l^i) = \frac{P(z_k|d_i) P(w_j|z_k) P(s_l^i|z_k, d_i, \boldsymbol{\alpha})}{\sum_m P(z_m|d_i) P(w_j|z_m) P(s_l^i|z_m, d_i, \boldsymbol{\alpha})} \tag{B.4}$$

B.1.2 Deriving the formula for updating the normalized inter-region relations r_{pl}^{ik}

Removing those terms that do not depend on s_l^i from the equation (eq. B.3), and expanding $P(s_l^i|z_k, d_i, \boldsymbol{\alpha})$ following equation (4.15), we obtain:

$$\log L_{spt} \geq \sum_{i,j,l,k} n^i(w_j, s_l^i) P(z_k|d_i, w_j, s_l^i) \log \sum_p \alpha_p^{ik} \lambda_{pl}^i \tag{B.5}$$

In order to apply again the Jensen's inequality, a normalize inter-region relation r_{pl}^{ik} that satisfy $\sum_{l=1}^{R_i} r_{pl}^{ik} = 1$ is defined, leading to this new expression for the lower bound of the L_{spt} log-likelihood:

$$\log L_{spt} \geq \sum_{i,j,l,k,p} n^i(w_j, s_l^i) P(z_k|d_i, w_j, s_l^i) r_{pl}^{ik} \log \frac{\alpha_p^{ik} \lambda_{pl}^i}{r_{pl}^{ik}} \tag{B.6}$$

Now, the maximization of this log-likelihood respect to r^{ik}_{pl} and subject to $\sum_{l=1}^{R_i} r^{ik}_{pl} = 1$ can be solved by using Langrange multipliers μ^{ik}_p as follows:

$$\log L_{r^{ik}_{pl}} = \log L_{spt} + \sum_{i,k,p} \mu^{ik}_p (\sum_l r^{ik}_{pl} - 1) \tag{B.7}$$

Taking derivatives with respect to r^{ik}_{pl}, we obtain:

$$\frac{\partial \log L_{r^{ik}_{pl}}}{\partial r^{ik}_{pl}} = \textstyle\sum_j n^i(w_j, s^i_l) P(z_k|d_i, w_j, s^i_l) \left[\log \frac{\alpha^{ik}_p \lambda^i_{pl}}{r^{ik}_{pl}} - 1\right] - \mu^{ik}_p \tag{B.8}$$

and setting the derivative to zero yields the updating equation for r^{ik}_{pl} (cf. eq. 4.21):

$$r^{ik}_{pl} = \frac{\alpha^{ik}_p \lambda^i_{pl}}{\sum\limits_{m=1}^{R_i} \alpha^{ik}_m \lambda^i_{ml}} \tag{B.9}$$

B.1.3 Deriving the formula for updating the importances $\boldsymbol{\alpha}^{ik}$

The update of the importances also requires a constrained maximization process, as stated in the next expression:

$$\log L_{\alpha^{ik}_p} = \log L_{spt} + \sum_{i,k} \mu_{ik} \left(\sum_l P(s^i_l|z_k, d_i, \boldsymbol{\alpha}) - 1 \right) \tag{B.10}$$

Using the expression in (4.15) and taking derivatives with respect to α^{ik}_p gives:

$$\frac{\partial \log L_{\alpha^{ik}_p}}{\partial \alpha^{ik}_p} = \sum_{j,l} n^i(w_j, s^i_l) P(z_k|d_i, w_j, s^i_l) r^{ik}_{pl} \frac{1}{\alpha^{ik}_p} - \mu_{ik} \chi^i_p \tag{B.11}$$

with χ^i_p as defined in eq. (4.20). Again, we obtain the updating equation for α^{ik}_p by setting this derivative to zero (cf. eq. 4.19):

$$\alpha^{ik}_p = \frac{\sum\limits_{j,l} n^i(w_j, s^i_l) P(z_k|d_i, w_j, s^i_l) r^{ik}_{pl}}{\chi^i_p \cdot \sum\limits_{j,l} n^i(w_j, s^i_l) P(z_k|d_i, w_j, s^i_l)} \tag{B.12}$$

B.2 Derivation of the formulas for the supervised RBLTM

This appendix provides the derivation of the updating equations of the inference process for the supervised version of RBLTM. Again, the index ranges are given here and omitted later on: $i = 1, 2, ..., D$, $j = 1, 2, ..., M$, $l, p = 1, 2, ..., R_i$, and $k, m = 1, 2, ..., K$.

B.2.1 Computation of a lower bound of the log-likelihood

As described in section 4.3.3, the logarithm of the posterior distribution $f(\theta)$ can be defined as follows :

$$\log f(\theta) = \log L + \log g_{img}(\theta) + \log g_{reg}(\theta) \tag{B.13}$$

where $\log L$ has been defined in eq. (B.2), and g_{img} and g_{reg} stand for the prior distributions of the image- and region-level parameters, respectively. The lower bound for $\log L$ can be obtained following the same procedure as in Appendix B.1.1.

B.2.2 Computation of the posterior probabilities of the latent variables for image-based annotations

In this section, we derive the extension of the unsupervised RBLTM to deal with image-level annotations. In order to obtain the update equation for the distribution of the topics given the document, we consider the terms of the posterior that depend on $P(z_k|d_i)$. In particular, using the prior distribution given in (4.23) and the lower bound of the log-likelihood given in (B.3), we obtain a lower bound of the distribution f_{img}:

$$\begin{aligned} \log f_{img} \geq & \sum_{i,j,l,k} n^i(w_j, s_l^i) P(z_k|d_i, w_j, s_l^i) \log P(z_k|d_i) \\ + & \sum_{i,k} (\beta_k^i - 1) P(z_k|d_i) + \sum_i \mu_i \left(\sum_k P(z_k|d_i) - 1 \right) \end{aligned} \tag{B.14}$$

where μ_i is a Lagrange multiplier resulting from the constrained optimization process. The derivative of this bound with respect to $P(z_k|d_i)$ gives:

$$\frac{\partial \log f_{img}}{\partial P(z_k|d_i)} = \sum_{j,l} n^i(w_j, s_l^i) P(z_k|d_i, w_j, s_l^i) \frac{1}{P(z_k|d_i)} + (\beta_k^i - 1) + \mu_i \quad \text{(B.15)}$$

Setting the derivative to zero and modeling the parameter $\beta_k^i = \epsilon_{IMG}^i \tilde{P}(z_k|d_i) + 1$ provides the update formula (cf. eq. 4.26):

$$P(z_k|d_i) = \frac{\sum_{j,l} n^i(w_j, s_l^i) P(z_k|d_i, w_j, s_l^i) + \epsilon_{IMG}^i \tilde{P}(z_k|d_i)}{\sum_{j,l} n^i(w_j, s_l^i) + \epsilon_{IMG}^i} \quad \text{(B.16)}$$

B.2.3 Deriving the formula for updating the importances $\boldsymbol{\alpha}^{ik}$ for region-based annotations

In this section, we derive the update equations of $\boldsymbol{\alpha}^{ik}$ when prior distributions are given at region-level. Again, a lower bound of the distribution f_{reg} is obtained by solving a constrained optimization problem:

$$\log f_{reg} \geq \log L_{spt} + \sum_{k,p} (\gamma_p^{ik} - 1) \log(\alpha_p^{ik} \chi_p^i) + \sum_{i,k} \mu_{ik} \left(\sum_l P(s_l^i|z_k, d_i, \boldsymbol{\alpha}) - 1 \right) \quad \text{(B.17)}$$

where we have made use of eq. (4.27). Computing the derivative with respect to α_p^{ik} gives:

$$\frac{\partial \log f_{reg}}{\partial \alpha_p^{ik}} = \sum_{j,l} n^i(w_j, s_l^i) P(z_k|d_i, w_j, s_l^i) r_{pl}^{ik} \frac{1}{\alpha_p^{ik}} + (\gamma_p^{ik} - 1) \frac{1}{\alpha_p^{ik}} - \mu_{ik} \chi_p^i \quad \text{(B.18)}$$

Setting the derivative to zero and modeling the parameter $\gamma_p^{ik} = \epsilon_{REG}^{ik} \tilde{\alpha}_p^{ik} + 1$

provides the update formula of the importances (cf. eq. 4.29):

$$\alpha_p^{ik} = \frac{\displaystyle\sum_{j,l} n^i(w_j, s_l^i) P(z_k|d_i, w_j, s_l^i) r_{pl}^{ik} + \epsilon_{REG}^{ik} \tilde{\alpha}_p^{ik}}{\chi_p^i \left[\displaystyle\sum_{j,l} n^i(w_j, s_l^i) P(z_k|d_i, w_j, s_l^i) + \epsilon_{REG}^{ik} \tilde{\Gamma}^{ik} \right]} \tag{B.19}$$

with $\tilde{\Gamma}^{ik}$ as defined in (4.30).

Appendix C

Derivation of the formulas for the RBLDA

C.1 Expansion of the lower bound

In this appendix we derive the equations of the RBLDA (see section 4.4).

As already mentioned in the document, a lower-bound was proposed over the posterior probability by introducing a variational distribution q:

$$\begin{aligned}\log p(\mathbf{h},\mathbf{v},\mathbf{g}|\Theta_p) \geq{} & E_q[\log p(\boldsymbol{\theta}|\boldsymbol{\alpha})] + \sum_{r=1}^{R_d}\bigg(E_q[\log p(z_r|\boldsymbol{\theta})] \\ & + E_q[\log p(h_r|z_r,\mathbf{a})] + E_q[\log p(l_r|z_r,\boldsymbol{\delta},\boldsymbol{\lambda})] + E_q[\log p(l_r|z_r,\boldsymbol{\beta})] \\ & + E_q[\log p(g_r|z_r,\mu,\Sigma)]\bigg) + \sum_{k=1}^{K} E_q[\log p(\boldsymbol{\delta}_k|\boldsymbol{\eta}_k)] + H(q)\end{aligned} \tag{C.1}$$

where $E_q[\cdot]$ denotes the expectation over the variational distribution q, and $H(\cdot)$ stands for the entropy of a distribution and obeys:

$$H(q) = -E_q[q(\boldsymbol{\theta}|\boldsymbol{\gamma})] - \sum_{r=1}^{R_d} E_q[\log q(z_r|\phi_r)] - \sum_{k=1}^{K} E_q[\log q(\delta_k|\chi_k)] \tag{C.2}$$

We next provide the expressions of each of them terms involved in the optimization:

$$E_q[\log p(\boldsymbol{\theta}|\boldsymbol{\alpha})] = \log\Gamma(\sum_j \alpha_j) - \sum_{k=1}^{K}\log\Gamma(\alpha_k) + \sum_{k=1}^{K}(\alpha_k - 1)\Big(\Psi(\gamma_k) - \Psi\big(\sum_j \gamma_j\big)\Big) \tag{C.3}$$

$$E_q[\log p(z_r|\boldsymbol{\theta})]) = \sum_{k=1}^{K}\phi_{rk}\Big(\Psi(\gamma_k) - \Psi\big(\sum_j \gamma_j\big)\Big) \tag{C.4}$$

$E_q[\log p(h_r|z_r, \mathbf{a})]$, to be derived in section C.3

$E_q[\log p(l_r|z_r, \boldsymbol{\delta}, \boldsymbol{\lambda})]$, to be derived in section C.2

$$E_q[\log p(l_r|z_r, \boldsymbol{\beta})] = \sum_{k,s}^{K,N_s}\phi_{rk}p\{l_r, s\}\beta_{ks} \tag{C.5}$$

$$E_q[\log p(g_r|z_r, \mu, \Sigma)] = \sum_{k=1}^{K}\phi_{rk}\Big[-\frac{K}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(g_r - \mu_k)^T\Sigma_k^{-1}(g_r - \mu_k)\Big] \tag{C.6}$$

$$E_q[\log p(\delta_k|\eta_k)] = \log\Gamma\big(\sum_j \eta_j\big) - \sum_{p=1}^{R_d}\log\Gamma(\eta_p) + \sum_p(\eta_p - 1)\Big(\Psi(\chi_{kp}) - \Psi(\sum_j \chi_{kj})\Big) \tag{C.7}$$

$$E_q[q(\boldsymbol{\theta}|\boldsymbol{\gamma})] = \log\Gamma\big(\sum_j \gamma_j\big) - \sum_{k=1}^{K}\log\Gamma(\gamma_k) + \sum_{k=1}^{K}(\gamma_k - 1)\Big(\Psi(\gamma_k) - \Psi\big(\sum_j \gamma_j\big)\Big) \tag{C.8}$$

$$E_q[\log q(z_r|\phi_r)] = \sum_{k=1}^{K}\phi_{rk}\log\phi_{rk} \tag{C.9}$$

$$E_q[\log q(\delta_k|\chi_k)] = \log\Gamma\big(\sum_j \chi_{kj}\big) - \sum_{p=1}^{R_d}\log\Gamma(\chi_{kp}) + \sum_{p=1}^{R_d}(\chi_{kp} - 1)\Big(\Psi(\chi_{kp}) - \Psi(\sum_j \chi_{kj})\Big) \tag{C.10}$$

where Ψ is the first derivative of the $\log \Gamma(\cdot)$ function and $p\{l_r, s\}$ represents the proportion of the region r that lies on the cell s of the grid (N_s cells).

C.2 Obtaining a lower bound of the context term

The term of the log-likelihood that is associated to a region context requires computing a lower bound in order to be tractable. Hence, if we introduce a new variational parameter $r_{tkpr} / \sum_{t=1}^{K} \sum_{p=1}^{R_d} r_{tkpr} = 1$, we can apply the Jensen's inequality and get the lower bound:

$$E_q[\log p(l_r|z_r, \boldsymbol{\delta}, \boldsymbol{\lambda})] = E_q\left[\log\left(\sum_{t=1}^{K}\sum_{p\neq l_r} c_{tz_r}\delta_{tp}\lambda_{pl_r}\right)\right] \geq E_q\left[\sum_{t,p\neq r}^{K} r_{tkpr}\log\frac{c_{tz_r}\delta_{tp}\lambda_{pl_r}}{r_{tkpr}}\right]$$

$$= \sum_{k,t,p\neq r}^{K,K} \phi_{rk} r_{tkpr}\left[\log\frac{c_{tk}\lambda_{pr}}{r_{tkpr}} + \Psi(\chi_{tp}) - \Psi\big(\sum_{m=1}^{R_d}\chi_{tm}\big)\right] \qquad \text{(C.11)}$$

C.3 Reducing the complexity of the appearance model

In this section, we reduce the complexity of the appearance model by computing a simplified lower bound of the term $E_q[\log p(h_r|z_r, \mathbf{a})]$. We follow the approach in [Jaakkola and Jordan, 2000], where the logistic function is symmetrized as:

$$\log f(x) = -log(1 + e^{-x}) = \frac{x}{2} - \log(e^{x/2} + e^{-x/2}) \qquad \text{(C.12)}$$

This gives the next expression:

$$E_q[\log p(h_r|z_r, \mathbf{a})] = \textstyle\sum_{r=1}^{R_d} \Bigg\{ E_q[\log n_k] + E_q\left[(1 - w_{z_r})z_r \frac{f_k(h_r)}{2}\right] -$$

$$-E_q\left[(1 - w_{z_r})z_r \frac{f_k(h_r)}{2}\right] - E_q\left[(1 - w_{z_r}z_r)\log(g_{kr})\right] - E_q\left[w_{z_r}\bar{z}_r \log(g_{kr})\right] \Bigg\} \text{(C.13)}$$

where $g_{kr} = e^{\frac{1}{2}f_k(h_r)} + e^{-\frac{1}{2}f_k(h_r)}$. Since $g(x) = -\log(e^{x/2} + e^{-x/2})$ is convex in the variable x^2, we can consider a tangent as a tight local lower bound of the function. This tangent is defined by the first order Taylor expansion in the variable x^2:

$$g(x) \geq g(\xi) + \frac{\partial g(\xi)}{\partial \xi^2}(x^2 - \xi^2) = -\frac{\xi}{2} + \log f(\xi) - \frac{1}{4\xi} \tanh(\xi/2)(x^2 - \xi^2) \quad \text{(C.14)}$$

This lower bound is exact when $\xi^2 = x^2$. Thus, the appearance term in the posterior now obeys:

$$E_q[\log p(h_r|z_r, \mathbf{a})] \geq \sum_{r=1}^{R_d} \sum_{k=1}^{K} \Bigg\{ \phi_{rk} \log n_k + (\phi_{rk} - w_k)\frac{f_k(h_r)}{2} + \Bigg[\phi_{rk}(1 - 2w_k) + w_k\Bigg] \cdot \\ \cdot \Bigg[-\frac{\xi}{2} - \log(1 + \exp(-\xi_{rk})) - \frac{1}{4\xi_{rk}} \tanh\left(\frac{\xi_{rk}}{2}\right)\left(f_k^2(h_r) - \xi_{rk}^2\right)\Bigg]\Bigg\} \quad \text{(C.15)}$$

where we have substituted the log function on f_{rk} by a quadratic function that yields a much simpler optimization. Furthermore, this approach requires the inclusion of a new variational parameter ξ.

C.4 Derivation of the formulas for the variational parameters

In this section, we provide the complete derivation of the update equations for the variational parameters to be computed in the E-step of the algorithm.

We start with those terms of the lower bound (LB) that depend on the appearance variational parameter ξ:

$$LB_\xi = \sum_{r,k} \Bigg[-\frac{\xi}{2} - log(1 + \exp(-\xi_{rk})) - \frac{1}{4\xi_{rk}} \tanh\left(\frac{\xi_{rk}}{2}\right)\left(f_k^2(h_r) - \xi_{rk}^2\right)\Bigg] \quad \text{(C.16)}$$

and computing the correspondent derivative gives:

$$\frac{\partial LB_\xi}{\partial \xi_{rk}} = -\lambda(\xi_{rk})(f_k^2(h_r) - \xi_{rk}^2) \quad \text{(C.17)}$$

where $\lambda(\xi_{rk}) = \frac{1}{4\xi_{rk}} \tanh\left(\frac{\xi_{rk}}{2}\right)$. Hence setting this derivative to zero provides the update equation:

$$\xi = \pm f_k(h_r) \quad \text{(C.18)}$$

Next we consider the terms of the lower-bound that depend on the normalized relations r_{kpr} and use some Lagrange multipliers μ_{kr} that ensure $\sum_{tp} r_{tkpr} = 1$:

$$\begin{aligned} LB_{\mathbf{r}} &= \sum_{t,k,r,p\neq r}^{K,K,R_d} \phi_{rk} r_{tkpr} \Big[\log \frac{c_{tk}\lambda_{pr}}{r_{tkpr}} + \Psi(\chi_{tp}) - \Psi\big(\sum_m \chi_{tm} \big) \Big] \\ &- \sum_{k,r}^{K,R_d} \mu_{kr} \Big(\sum_{t,p}^{K,R_d} r_{tkpr} - 1 \Big) \end{aligned} \tag{C.19}$$

We now compute the derivative of this expression with respect to $\mathbf{r}$ and add a Lagrange multiplier that ensures what gives:

$$\frac{\partial LB_r}{\partial r_{tkpr}} = \phi_{rk} \Big(\log \frac{c_{tk}\lambda_{pr}}{r_{tkpr}} + \Psi(\chi_{tp}) - \Psi\big(\sum_m \chi_{tm} \big) - 1 \Big) - \mu_{kr} \tag{C.20}$$

And setting its derivative to zero gives the final update equation for the normalized relationships:

$$r_{tkpr} \propto c_{tk}\lambda_{pr} \exp\big(\Psi(\chi_{kp}) \big) \tag{C.21}$$

We now work on the expression that depends on the parameter $\boldsymbol{\chi}$:

$$\begin{aligned} LB_{\boldsymbol{\chi}} = &\sum_{k,p}^{K,R_d} \Big(\Psi(\chi_{kp}) - \Psi\big(\sum_m \chi_{km} \big) \Big) \Big[\eta_p - \chi_{kp} + \sum_{t,r\neq p}^{K} \phi_{rk} r_{tkpr} \Big] \\ &- \sum_{k=1}^{K} \log\Gamma\big(\sum_{m=1}^{R_d} \chi_{km} \big) + \sum_{k,p}^{K,R_d} \log\Gamma(\chi_{kp}) \end{aligned} \tag{C.22}$$

Computing the derivative with respect to χ_{kp} gives:

$$\begin{aligned} \frac{\partial LB_{\boldsymbol{\chi}}}{\partial \chi_{kp}} = &\Psi'(\chi_{kp}) \Big[\eta_p - \chi_{kp} + \sum_{t,r\neq p}^{K} \phi_{rk} r_{tkpr} \Big] \\ &- \Psi'\Big(\sum_m \chi_{km} \Big) \Big[\sum_{m=1}^{R_d} \Big(\eta_m - \chi_{km} + \sum_{t,r\neq m}^{K} \phi_{rk} r_{tkmr} \Big) \Big] \end{aligned} \tag{C.23}$$

Thus, if we set this derivative to zero for every value of m we get the update equation for the new variational parameter as:

$$\chi_{kp} = \eta_p + \sum_{t,r\neq p}^{K} \phi_{rk} r_{tkpr} \tag{C.24}$$

Finally, we consider the lower bound that depends on the variational multinomial $\boldsymbol{\phi}$ and add a Lagrange parameter τ so that $\sum_k \phi_{rk} = 1$:

$$\begin{aligned} LB_{\boldsymbol{\phi}} = \sum_{k,r}^{K,R_d} \phi_{rk} \Big\{ & \Psi(\gamma_k) - \Psi\big(\sum_j \gamma_j\big) + \log n_k - w_k \xi_{rk} + (2w_k - 1)\log(1 + \exp(-\xi_{rk})) \\ & + \sum_{t,p \neq r}^{K} r_{tkpr} \left[\log \frac{c_{tk}\lambda_{pr}}{r_{tkpr}} + \Psi(\chi_{tp}) - \Psi\big(\sum_{m=1}^{R_d} \chi_{tm}\big) \right] + \sum_s^{N_s} p\{l_r, s\}\beta_{ks} - \frac{K}{2}\log 2\pi \\ & - \frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(g_r - \mu_k)^T \Sigma^{-1}(g_r - \mu_k) - \log \phi_{rk} \Big\} - \sum_{r=1}^{R_d} \tau_r \Big(\sum_{k=1}^{K} \phi_{rk} - 1 \Big) \end{aligned} \quad \text{(C.25)}$$

The derivative with respect to ϕ_{rk} is:

$$\begin{aligned} \frac{\partial LB_{\boldsymbol{\phi}}}{\partial \phi_{rk}} = & \Psi(\gamma_k) - \Psi\big(\sum_j \gamma_j\big) + \log n_k - w_k \xi_{rk} + (2w_k - 1)\log(1 + \exp(-\xi_{rk})) \\ & + \sum_{t,p \neq r}^{K} r_{tkpr} \left[\log \frac{c_{tk}\lambda_{pr}}{r_{tkpr}} + \Psi(\chi_{tp}) - \Psi\big(\sum_{m=1}^{R_d} \chi_{tm}\big) \right] + \sum_s^{N_s} p\{l_r, s\}\beta_{ks} - \frac{K}{2}\log 2\pi \\ & - \frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(g_r - \mu_k)^T \Sigma^{-1}(g_r - \mu_k) - \log \phi_{rk} - 1 - \tau_r \end{aligned} \quad \text{(C.26)}$$

And setting this derivative to zero gives the update expression of the multinomial:

$$\begin{aligned} \phi_{rk} \propto \exp \Bigg\{ & \Psi(\gamma_k) + \log n_k - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(g_k - \mu_k)\Sigma^{-1}(g_k - \mu_k) + w_k \xi_{rk} \qquad \text{(C.27)} \\ & + (2w_k - 1)\log(1 + \exp(-\xi_{rk})) + \sum_{t=1}^{K} \sum_{p \neq r} \left[\log \frac{c_{tk}\lambda_{pr}}{r_{tkpr}} + \Psi(\chi_{tp}) - \Psi\Big(\sum_{m=1}^{R_d} \chi_{tm} \Big) \right] \Bigg\} \end{aligned}$$

It is noteworthy that w_k has a value in a supervised training environment in order to learn the negative samples whereas it is set to zero in test and unsupervised environments. Additionally, equations for the variational Dirichlet $\boldsymbol{\gamma}$ and the Dirichlet parameters $\boldsymbol{\alpha}$ are not included since they do not differ from the original LDA proposal (see [Blei et al., 2003]).

C.5 Derivation of the formulas for the model parameters

In this section we provide complete derivations of the update equations of the model parameters that are computed in the M-step of the inference algorithm. Since some of the parameters are corpus-dependent, a new index d points to each document in the corpus.

We start by computing the derivative of the eq. (C.18) with respect to c_{tk} and add a Lagrange multiplier that ensures $\sum_t c_{tk} = 1$ and obtain:

$$\frac{\partial LB_c}{\partial c_{tk}} = \sum_{d,r,p\neq r}^{D,R_d} \phi_{drk} r_{tkpr} \frac{1}{c_{tk}} - \mu_k \tag{C.28}$$

Note the dependence on the document index d, since $\mathbf{c}$ is a corpus-based variable. We finish with the update equation for $\mathbf{c}$:

$$c_{tk} \propto \sum_{d,r,p\neq r}^{D,R_d} \phi_{dnk} r_{tkpr} \tag{C.29}$$

Now, we consider terms of the Lower Bound that depend on the Kernel Logistic Regressor (eq. (C.14)) and introduce a L2 regularization element the following expression needs to be maximized:

$$LB_{f_k} = \sum_{r=1}^{R_d} \sum_{k=1}^{K} C_{rk}^{(1)} f_{rk} - C_{rk}^{(2)} f_k^2(h_r) - \frac{\mu}{2} \|f\|_{\mathcal{H}_k}^2 \tag{C.30}$$

where $\mathcal{H}$ stands for the Reproducing Kernel Hilbert Space (RKHS), and the parameters $C^{(1)}, C^{(2)}$ are:

$$C_{rk}^{(1)} = \frac{1}{2}(\phi_{rk} - w_k) \tag{C.31}$$

$$C_{rk}^{(2)} = [\phi_{rk}(1 - 2w_k) + w_k] \frac{1}{4\xi_{rk}} \tanh\left(\frac{\xi_{rk}}{2}\right) \tag{C.32}$$

Thus, in order to obtain the optimal parameters of the regressors $\mathbf{a}_k$, we can use an iterative Newton-Raphson method so that, at iteration t:

$$\mathbf{a}_k^{(t+1)} = \mathbf{a}_k^{(t)} - H_k^{-1} \nabla_k \tag{C.33}$$

Using the dual form proposed in eq. (4.36), the values of the gradient ∇_k and the Hessian H_k obey:

$$\nabla_k = K_k^T C^{(1)} - 2K_k^T(C^{(2)} \cdot f_k) - \frac{\mu}{2} K'_k \mathbf{a}_k \tag{C.34}$$

$$H_k = -2K_k^T \operatorname{diag}(C^{(2)}) K_k - \frac{\mu}{2} K'_k \tag{C.35}$$

where K and K' stand for the data Kernel matrix and the regularization matrix, respectively, and $\cdot$ represents the Hadamard product (element wise) between two matrices.

Furthermore, we next show the terms of the log-likelihood that depend on the normalization factor n_k and add a Lagrange parameter that ensures that the appearance distribution is a probability density function along the potential values of the histogram (following the aforementioned approximation in which the normalization is provided over the training set):

$$LB_{n_k} = \sum_{rk} \phi_{rk} \log n_k - \sum_{k=1}^{K} \mu_k \left(\sum_{p=1}^{R_d} \frac{n_k}{1+e^{-f_{kp}}} - 1 \right) \tag{C.36}$$

Computing the derivative of this expression with respect to n_k gives:

$$\frac{\partial LB_n}{\partial n_k} = \sum_r \phi_{rk} - \mu_k \sum_{p=1}^{R_d} \frac{n_k}{1+e^{-f_{kp}}} \tag{C.37}$$

Finally, setting this derivative to zero gives the final expression for the normalization term:

$$n_k^{-1} = \sum_{r=1}^{R_d} \frac{1}{1+\exp(-f_k(h_r))} \tag{C.38}$$

We finish with the means and covariances of the Gaussian estimates on the global probabilities. The terms of the LB that incorporate both terms are:

$$LB_{\{\mu_k,\Sigma_k\}} = \sum_{d,r,k}^{D,R_d,K} \phi_{drk} \left[-\frac{K}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(g_{dr}-\mu_k)^T \Sigma_k^{-1}(g_{dr}-\mu_k) \right] \tag{C.39}$$

Computing the derivative with respect to μ_k gives:

$$\frac{\partial LB_{\{\mu_k,\Sigma_k\}}}{\partial \mu_k} = \sum_{d,r}^{D,R_d} \phi_{drk} \left[\Sigma_k^{-1}(g_{dr} - \mu_k) \right] \tag{C.40}$$

So that the update equations obeys:

$$\mu_k = \frac{\sum_{d,r}^{D,R_d} \phi_{drk} g_{dr}}{\sum_{d,r}^{D,R_d} \phi_{drk}} \tag{C.41}$$

Now, the derivative with respect to the covariance matrix Σ_k is:

$$\frac{\partial LB_{\{\mu_k,\Sigma_k\}}}{\partial \Sigma_k} = -\sum_{d,r}^{D,R_d} \frac{1}{2} \phi_{drk} \left[\Sigma_k^{-1} + \Sigma_k^{-1}(g_{dr} - \mu_k)(g_{dr} - \mu_k)^T \Sigma_k^{-1} \right] \tag{C.42}$$

which gives the final update for the covariance:

$$\Sigma_k = \frac{\displaystyle\sum_{d,r}^{D,R_d} \phi_{drk}(g_{dr} - \mu_k)(g_{dr} - \mu_k)^T}{\displaystyle\sum_{d,r}^{D,R_d} \phi_{drk}} \tag{C.43}$$

Bibliography

[Ano, 2008] (2008). Learning hybrid models for image annotation with partially labeled data. In *Annual Conference on Neural Information Processing Systems (NIPS 08)*. 54

[Adamek and O'Connor., 2007] Adamek, T. and O'Connor., N. (2007). Using dempster-shafer theory to fuse multiple information sources in region-based segmentation. In *Proceedings of the 14th IEEE International Conference on Image Processing*. 17, 36

[Adamek and O'Connor, 2007] Adamek, T. and O'Connor, N. (2007). Using Dempster-Shafer Theory to Fuse Multiple Information Sources in Region-Based Segmentation. In *ICIP 2007 - Proceedings of the 14th IEEE International Conference on Image Processing*. 19, 30

[Aghbari et al., 1998] Aghbari, Z., Kaneko, K., and Makinouchi, A. (25-28 Aug 1998). A motion-location based indexing method for retrieving mpeg videos. *Database and Expert Systems Applications, 1998. Proceedings. Ninth International Workshop on*, pages 102–107. 19

[Bay et al., 2008] Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. (2008). Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110:346–359. 50

[Bennstrom and Casas, 2004] Bennstrom, C. and Casas, J. (14-16 July 2004). Binary-partition-tree creation using a quasi-inclusion criterion. *Information Visu-*

alisation, 2004. IV 2004. Proceedings. Eighth International Conference on, pages 259–264. 19

[Blei and Jordan, 2003] Blei, D. M. and Jordan, M. I. (2003). Modeling annotated data. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 127–134, New York, NY, USA. ACM. 62, 63

[Blei and Mcauliffe, 2007] Blei, D. M. and Mcauliffe, J. D. (2007). Supervised topic models. 54

[Blei et al., 2003] Blei, D. M., Ng, A. Y., Jordan, M. I., and Lafferty, J. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:2003. viii, 12, 44, 45, 53, 54, 59, 60, 96, 98, 148

[Boix et al., 2010] Boix, X., Gonfaus, J. M., de Weijer, J. V., Bagdanov, A. D., Serrat, J., , and Gonzalez, J. (2010). Harmony potentials: Fusing global and local scale for semantic image segmentation. *International Journal of Computer Vision (accepted)*. 119

[Bosch et al., 2007] Bosch, A., Zisserman, A., and Munoz, X. (2007). Representing shape with a spatial pyramid kernel. In *CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 401–408, New York, NY, USA. ACM. 44

[Bosch et al., 2008] Bosch, A., Zisserman, A., and Muoz, X. (2008). Scene classification using a hybrid generative/discriminative approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(4):712–727. 61

[Cao and Fei-Fei, 2007] Cao, L. and Fei-Fei, L. (2007). Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. *IEEE 11th International Conference on Computer Vision, 2007. ICCV 2007*, pages 1–8. 61, 87, 100, 101

[Chang and Lin, 2001] Chang, C. C. and Lin, C. J. (2001). *LIBSVM: a library for support vector machines*. 100, 101

[Chen et al., 2010] Chen, Q., Song, Z., Liu, S., Chen, X., Yuan, X., Chua, T.-S., Yan, S., Hua, Y., Huang, Z., and Shen, S. (2010). Boosting classification with exclusive context. 118, 120

[Choi et al., 1997] Choi, J. G., Lee, S.-W., and Kim, S.-D. (1997). Spatio-temporal video segmentation using a joint similarity measure. *IEEE Transactions on Circuits and Systems for Video Technology*, 7(2):279–286. 17, 19

[Csurka et al., 2004] Csurka, G., Dance, C. R., Fan, L., Willamowski, J., and Bray, C. (2004). Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22. viii, 47, 52

[Dalal and Triggs, 2005] Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, pages 886–893, Washington, DC, USA. IEEE Computer Society. 50, 118

[Davies, 2005] Davies, E. R. (2005). *Machine Vision, Third Edition: Theory, Algorithms, Practicalities (Signal Processing and its Applications)*. Morgan Kaufmann, 3 edition. 1, 2

[del Blanco et al., 2007] del Blanco, C. R., Jaureguizar, F., Salgado, L., and Garcia, N. (Sept. 16 2007-Oct. 19 2007). Target detection through robust motion segmentation and tracking restrictions in aerial flir images. *IEEE International Conference on Image Processing, 2007. ICIP 2007.*, 5:V–445–V–448. 23

[Duda et al., 2001] Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. Wiley, New York, 2. edition. 51

[Eriksson et al., 2007] Eriksson, A., Olsson, C., and Kahl, F. (2007). Normalized cuts revisited: A reformulation for segmentation with linear grouping constraints. In *ICCV 2007. IEEE 11th International Conference on Computer Vision, 2007.*, pages 1–8. 17

[Everingham et al., 2009] Everingham, M., Gool, L. V., Williams, C. K. I., Winn, J., and Zisserman, A. (2009). The PASCAL Visual Object Classes Challenge 2009 (VOC2009). http://www.pascal-network.org/challenges/VOC/voc2009/. 52, 102

[Everingham et al., 2010] Everingham, M., Gool, L. V., Williams, C. K. I., Winn, J., and Zisserman, A. (2010). The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html. x, xxvi, 13, 84, 85, 99

[Fei-Fei and Perona, 2005] Fei-Fei, L. and Perona, P. (2005). A bayesian hierarchical model for learning natural scene categories. *International Conferece in Computer Vision and Pattern Recognition*, pages 524–531. 47

[Felzenszwalb et al., 2010a] Felzenszwalb, P. F., Girshick, R. B., and McAllester, D. A. (2010a). Cascade object detection with deformable part models. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pages 2241–2248. xiii, 127

[Felzenszwalb et al., 2010b] Felzenszwalb, P. F., Girshick, R. B., McAllester, D. A., and Ramanan, D. (2010b). Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1627–1645. xiii, 127

[Felzenszwalb and Huttenlocher, 2004] Felzenszwalb, P. F. and Huttenlocher, D. P. (2004). Efficient graph-based image segmentation. *Int. J. Comput. Vision*, 59:167–181. 64

[Fergus et al., 2005] Fergus, R., Fei-Fei, L., Perona, P., and Zisserman, A. (2005). Learning object categories from google's image search. volume 2, pages 1816–1823 Vol. 2. 60

[Galleguillos et al., 2008] Galleguillos, C., Rabinovich, A., and Belongie, S. (2008). Object categorization using co-occurrence, location and appearance. pages 1–8. 45

[Ge et al., 2006] Ge, F., Wang, S., and Liu, T. (17-22 June 2006). Image-segmentation evaluation from the perspective of salient object extraction. *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, 1:1146–1153. 35

[Goldberger and Greenspan, 2006] Goldberger, J. and Greenspan, H. (March 2006). Context-based segmentation of image sequences. *Transactions on Pattern Analysis and Machine Intelligence*, 28(3):463–468. 15, 18, 25, 26

[Gonzalez-Diaz et al., 2007] Gonzalez-Diaz, I., de Frutos-Lopez, M., Sanz-Rodriguez, S., and de Maria, D. (2007). Adaptive multi-pattern fast block-matching algorithm based on motion classification techniques. In *IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. ICASSP 2007.*, volume 1, pages I–1177–I–1180. 18

[Gonzalez-Diaz and de Maria, 2007] Gonzalez-Diaz, I. and de Maria, F. D. (2007). Improved motion classification techniques for adaptive multi-pattern fast block-matching algorithm. In *IEEE International Conference on Image Processing, 2007. ICIP 2007.*, volume 2, pages II–485–II–488. 18

[Gonzalez-Diaz and de Maria, 2008] Gonzalez-Diaz, I. and de Maria, F. D. (2008). Adaptive multipattern fast block-matching algorithm based on motion classification techniques. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(10):1369–1382. 18

[González-Díaz and de María, 2011] González-Díaz, I. and de María, F. D. (2011). Improving the appearance model in a spatially-aware latent topic model for image representation and segmentation. In *13th International Conference on Computer Vision, 2011. (ICCV'11). [submitted].* 79

[González-Díaz et al., 2009a] González-Díaz, I., García-García, D., and de María., F. D. (2009a). A spatially-aware generative model for image classification. In *International Conference on Image Processing, 2009. ICIP '09.* 63

[González-Díaz et al., 2009b] González-Díaz, I., Gómez-Verdejo, V., Martínez-Ramon, M., de María, F. D., and Arenas-García, J. (2009b). Uc3m at trecvid 2009. In *2009 TRECVID Workshop.* xiii, 41, 127

[González-Díaz et al., 2008] González-Díaz, I., McGuinness, K., Adamek, T., O'Connor, N., and de María, F. D. (2008). Incorporating spatio-temporal mid-level features in a region segmentation algorithm for video sequences. In *IEEE International Conference on Image Processing, 2008. (ICIP'08).* 20

[Gould et al., 2009a] Gould, S., Fulton, R., and Koller, D. (2009a). Decomposing a scene into geometric and semantically consistent regions. In *IEEE 12th International Conference on Computer Vision, 2009*, pages 1–8. 45

[Gould et al., 2009b] Gould, S., Gao, T., and Koller, D. (2009b). Region-based segmentation and object detection. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. K. I., and Culotta, A., editors, *Advances in Neural Information Processing Systems 22*, pages 655–663. 45

[Grauman and Darrell, 2005] Grauman, K. and Darrell, T. (2005). The pyramid match kernel: discriminative classification with sets of image features. In *IEEE International Conference on Computer Vision*, volume 2, pages 1458–1465 Vol. 2. 44, 53

[Green and Yandell, 1985] Green, P. and Yandell, B. (1985). Semi-parametric generalized linear models. Technical Report 2847, University of Wisconsin-Madison. 88

[Greenspan et al., 2004] Greenspan, H., Goldberger, J., and Mayer, A. (Mar 2004). Probabilistic space-time video modeling via piecewise GMM. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(3):384–396. 16, 18, 20, 36

[Greenspan et al., 2006] Greenspan, H., Ruf, A., and Goldberger, J. (2006). Constrained gaussian mixture model framework for automatic segmentation of mr brain images. *IEEE Transactions on Medical Imaging*, 25(9):1233–1245. 17

[Griffin et al., 2007] Griffin, G., Holub, A., and Perona, P. (2007). Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology. 52

[Haralick and Shapiro, 1992] Haralick, R. and Shapiro, L. (1992). *Computer and Robot Vision, Volume 1.* Addison Wesley. 48

[Harris and Stephens, 1988] Harris, C. and Stephens, M. (1988). A combined corner and edge detector. In *Proceedings of the 4th Alvey Vision Conference*, pages 147–151. 48

[Hastie and Tibshirani, 1987] Hastie, T. and Tibshirani, R. (1987). Generalized additive models: Some applications. *Journal of the American Statistical Association*, 82:371–386. 88

[Hofmann, 2001] Hofmann, T. (2001). Unsupervised learning by Probabilistic Latent Semantic Analysis. *Machine Learning*, 42(1/2):177–196. viii, 12, 44, 45, 53, 54, 56

[Hou et al., 2010] Hou, Y., Sun, X., Lun, X., and Lan, J. (2010). Gaussian mixture model segmentation algorithm for remote sensing image. In *2010 International Conference on Machine Vision and Human-Machine Interface (MVHI)*, pages 275—-278. 17

[Hu and Li, 2010] Hu, X. and Li, W. (2010). Multi-scale optical flow estimation of the video based on gradient optimization. In *3rd International Congress on Image and Signal Processing (CISP), 2010*, volume 1, pages 335–339. 18

[Huang and Dom, 1995] Huang, Q. and Dom, B. (1995). Quantitative methods of evaluating image segmentation. In *Proceedings of International Conference on Image Processing, 1995*, volume 3, pages 53–56 vol.3. 34

[Ince and Konrad, 2008] Ince, S. and Konrad, J. (2008). Occlusion-aware optical flow estimation. *IEEE Transactions on Image Processing*, 17(8):1443–1451. 18

[Jaakkola and Jordan, 2000] Jaakkola, T. S. and Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10:25–37. 94, 145

[Jain and Dubes, 1988] Jain, A. and Dubes, R. (1988). *Algorithms for Clustering Data*. Prentice Hall. v, 9

[Jain et al., 2004] Jain, A., Topchy, A., Law, M., and Buhmann, J. (2004). Landscape of clustering algorithms. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 1, pages 260–263 Vol.1. v, 9

[Ji et al., 2010] Ji, R., Yao, H., Sun, X., Zhong, B., and Gao, W. (2010). Towards semantic embedding in visual vocabulary. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 918–925. 51

[Jordan et al., 1999] Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. (1999). An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233. ix, 59

[Kohonen, 1997] Kohonen, T., editor (1997). *Self-organizing maps*. Springer-Verlag New York, Inc., Secaucus, NJ, USA. 51

[Kullback and Leibler, 1951] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22:49–86. x

[Kwok and Constantinides, 1997] Kwok, S. and Constantinides, A. (Feb 1997). A fast recursive shortest spanning tree for image segmentation and edge detection. *IEEE Transactions on Image Processing*, 6(2):328–332. 16, 17, 29

[Lafferty et al., 2004] Lafferty, J., Zhu, X., and Liu, Y. (2004). Kernel conditional random fields: representation and clique selection. In *Proceedings of the twenty-first international conference on Machine learning*, ICML '04, pages 64–, New York, NY, USA. ACM. 91

[Larlus et al., 2010] Larlus, D., Verbeek, J., and Jurie, F. (2010). Category level object segmentation by combining bag-of-words models with dirichlet processes and random fields. *International Journal of Computer Vision*, 88(2):238–253. 45

[Lazebnik et al., 2006] Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2169–2178. 44, 114

[Lee and Grauman, 2010] Lee, Y. and Grauman, K. (2010). Collect-cut: Segmentation with top-down cues discovered in multi-object images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 45

[Lehmann, 2011] Lehmann, F. (2011). Turbo segmentation of textured images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(1):16–29. 18

[Lewis, 1998] Lewis, D. D. (1998). Naive (bayes) at forty: The independence assumption in information retrieval. In *Proceedings of the 10th European Conference on Machine Learning*, pages 4–15, London, UK. Springer-Verlag. 53

[Li et al., 2009] Li, L.-J., Socher, R., and Fei-Fei, L. (2009). Towards total scene understanding:classification, annotation and segmentation in an automatic framework. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 61, 63

[Lievin and Luthon, 2004] Lievin, M. and Luthon, F. (2004). Nonlinear color space and spatiotemporal mrf for hierarchical segmentation of face features in video. *IEEE Transactions on Image Processing*, 13(1):63–71. 18

[Liu and Chen, 2006] Liu, D. and Chen, T. (2006). Semantic-shift for unsupervised object detection. In *Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop*, page 16, Washington, DC, USA. IEEE Computer Society. 60

[Lowe, 2004] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110. 48, 49, 50, 66

[Marimon et al., 2010] Marimon, D., Bonnin, A., Adamek, T., and Gimeno, R. (2010). Darts: Efficient scale-space extraction of daisy keypoints. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010*, pages 2416–2423. 50

[Marr, 1982] Marr, D. (1982). *Vision: a computational investigation into the human representation and processing of visual information.* W. H. Freeman, San Francisco. 1

[Martin et al., 2001] Martin, D., Fowlkes, C., Tal, D., and Malik, J. (2001). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. 8th Int'l Conf. Computer Vision*, volume 2, pages 416–423. 34

[Matas et al., 2002] Matas, J., Chum, O., Urban, M., and Pajdla, T. (2002). Robust wide baseline stereo from maximally stable extremal. In *In British Machine Vision Conference*, pages 384–393. 48

[Matusita, 1955] Matusita, K. (1955). Decision rules, based on the distance, for problems of fit, two samples, and estimation. *Annals of Mathematical Statisticals*, 26(4):631–640. 29

[McGuinness et al., 2006] McGuinness, K., Keenan, G., Adamek, T., and O'Connor, N. (2006). A framework for integrating and evaluating automatic region-based segmentation algorithms. In *SAMT 2006 - Poster and Demo Proceedings of The First International Conference on Semantics And Digital Media Technology*, pages 41–42. 34

[McGuinness et al., 2007] McGuinness, K., Keenan, G., Adamek, T., and O'Connor, N. (2007). Image segmentation evaluation using an integrated framework. In *VIE 2007 - Proceedings of the IET 4th International Conference on Visual Information Engineering 2007*. 34

[Mikolajczyk and Schmid, 2002] Mikolajczyk, K. and Schmid, C. (2002). An affine invariant interest point detector. In *Proceedings of the 7th European Conference on Computer Vision-Part I*, ECCV '02, pages 128–142, London, UK, UK. Springer-Verlag. 48

[Mikolajczyk et al., 2005] Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., and Van Gool, L. (2005). A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1/2):43–72. 48

[Moscheni et al., 1998] Moscheni, F., Bhattacharjee, S., and Kunt, M. (1998). Spatiotemporal segmentation based on region merging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:897–915. 17

[National Institute of Standards and Technology, 2006] National Institute of Standards and Technology, N. (2006). Guidelines for the trecvid 2006 evaluation. Published online: http://www-nlpir.nist.gov/projects/tv2006/tv2006.html. viii, 33

[Nister and Stewenius, 2006] Nister, D. and Stewenius, H. (2006). Scalable recognition with a vocabulary tree. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006*, volume 2, pages 2161–2168. 51

[Ojala et al., 1996] Ojala, T., Pietikainen, M., and Harwood, D. (1996). A comparative study of texture measures with classification based on feature distributions. 29(1):51–59. 118

[Oliva and Torralba, 2001] Oliva, A. and Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145–175. 118

[Philbin et al., 2008] Philbin, J., Chum, O., Isard, M., Sivic, J., and Zisserman, A. (2008). Lost in quantization: Improving particular object retrieval in large scale image databases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 52

[Rosenblatt, 1962] Rosenblatt, F. (1962). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan Books. 31

[Russell et al., 2006] Russell, B., Freeman, W., Efros, A., Sivic, J., and Zisserman, A. (2006). Using multiple segmentations to discover objects and their extent in image collections. volume 2, pages 1605–1614. 61

[Schölkopf and Smola, 2002] Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels*. The MIT Press, Cambridge, MA. 53

[Shi and Malik, 2000] Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905. 17

[Shotton et al., 2009] Shotton, J., Winn, J., Rother, C., and Criminisi, A. (2009). Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *Int. J. Comput. Vision*, 81(1):2–23. 45, 87

[Sivic et al., 2005] Sivic, J., Russell, B. C., Efros, A. A., Zisserman, A., and Freeman, W. T. (2005). Discovering objects and their location in images. In *IEEE International Conference on Computer Vision*, volume 1, pages 370–377. 47, 60, 101

[Sivic and Zisserman, 2003] Sivic, J. and Zisserman, A. (2003). Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the International Conference on Computer Vision*, volume 2, pages 1470–1477. viii, 47

[Smeaton et al., 2009] Smeaton, A. F., Over, P., and Kraaij, W. (2009). High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements. In Divakaran, A., editor, *Multimedia Content Analysis, Theory and Applications*, pages 151–174. Springer Verlag, Berlin. 52

[Smeulders et al., 2000] 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. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22:1349–1380. 15

[Smola and Schökopf, 2000] Smola, A. J. and Schökopf, B. (2000). Sparse greedy matrix approximation for machine learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 911–918, San Francisco, CA,greenspan USA. Morgan Kaufmann Publishers Inc. 91

[Snoek and Worring, 2009] Snoek, C. G. M. and Worring, M. (2009). Concept-based video retrieval. *Foundations and Trends in Information Retrieval*, 4(2):215–322. 47

[Stewart, 1999] Stewart, C. V. (1999). Robust parameter estimation in computer vision. *SIAM Rev.*, 41:513–537. 23

[Sudderth et al., 2007] Sudderth, E. B., Torralba, A., Freeman, W. T., and Willsky, A. S. (2007). Describing visual scenes using transformed objects and parts. *International Journal of Computer Vision*. 61

[Sun et al., 2010] Sun, T., Jiang, X., Fu, G., Li, R., Feng, B., Wang, S., Sun, T., and Jiang, X. (2010). Image semantic recognition scheme with semantic-binding hierarchical visual vocabulary model. In *Image and Signal Processing (CISP), 2010 3rd International Congress on*, volume 4, pages 1576–1581. 51

[Szeliski, 2011] Szeliski, R. (2011). *Computer Vision: Algorithms and Applications*. Springer. 6, 8, 9

[Takacs and Demiris, 2008] Takacs, B. and Demiris, Y. (2008). Balancing spectral clustering for segmenting spatio-temporal observations of multi-agent systems. In *Eighth IEEE International Conference on Data Mining, 2008. ICDM '08.*, pages 580–587. 17

[Teh et al., 2006] Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581. xiii, 127

[Titterington et al., 1985] Titterington, D., Smith, A., and Makov, U. (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons. vii

[Tsaig and Averbuch, 2001] Tsaig, Y. and Averbuch, A. (2001). A region-based mrf model for unsupervised segmentation of moving objects in image sequences. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001. CVPR 2001.*, volume 1, pages I–889–I–896 vol.1. 18

[Tuytelaars et al., 2010] Tuytelaars, T., Lampert, C., Blaschko, M., and Buntine, W. (2010). Unsupervised object discovery: A comparison. *International Journal of Computer Vision*, 88:284–302. 46

[Tuytelaars and Mikolajczyk, 2008] Tuytelaars, T. and Mikolajczyk, K. (2008). Local invariant feature detectors: a survey. *Found. Trends. Comput. Graph. Vis.*, 3:177–280. 48

[van Gemert et al., 2010] van Gemert, J. C., Snoek, C. G. M., Veenman, C. J., Smeulders, A. W. M., and Geusebroek, J. M. (2010). Comparing compact codebooks for visual categorization. *Computer Vision and Image Understanding*, 114(4):450–462. 51

[Varma and Ray, 2007] Varma, M. and Ray, D. (2007). Learning the discriminative power-invariance trade-off. pages 1–8. 44

[Wahba et al., 1993] Wahba, G., Gu, C., and Wang, Y. (1993). Soft classification, a.k.a. risk estimation, via penalized log likelihood and smoothing spline analysis of variance. In *The Mathematics of Generalization*. Addison-Wesley. 88

[Wallraven et al., 2003] Wallraven, C., Caputo, B., and Graf, A. (2003). Recognition with local features: the kernel recipe. In *IEEE International Conference on Computer Vision*, pages 257–264 vol.1. 44, 53

[Wang et al., 2009] Wang, C., Blei, D., and Fei-Fei, L. (2009). Simultaneous image classification and annotation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 61, 63, 87

[Wang and Grimson, 2007] Wang, X. and Grimson, E. (2007). Spatial latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, volume 20. 61

[Wang et al., 2005] Wang, Y., Loe, K., Tan, T., and Wu, J. (July 2005). Spatiotemporal video segmentation based on graphical models. *IEEE Transactions on Image Processing*, 14(7):937–947. 16

[Williams and Seeger, 2001] Williams, C. and Seeger, M. (2001). Using the nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems 13*, pages 682–688. MIT Press. 91

[Zhang and Zhang, 2004] Zhang, R. and Zhang, Z. M. (2004). Hidden semantic concept discovery in region based image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.* 61, 87

[Zhao et al., 2010] Zhao, B., Fei-Fei, L., and Xing, E. P. (2010). Image segmentation with topic random field. In *Proceedings of the 11th European conference on Computer vision: Part V*, ECCV'10, pages 785–798, Berlin, Heidelberg. Springer-Verlag. 62, 101

[Zhu and Hastie, 2001] Zhu, J. and Hastie, T. (2001). Kernel logistic regression and the import vector machine. In *Journal of Computational and Graphical Statistics*, pages 1081–1088. MIT Press. 89, 91