A Higher-level Visual Representation for Semantic Learning in Image Databases

Ismail El Sayad

T H È S E

pour obtenir le titre de

Docteur

de l'Université Lille 1 Sciences et Technologies

Mention : Informatique

Présentée et soutenue par

Ismail El Sayad

Une représentation visuelle avancée pour l'apprentissage sémantique dans les bases d'images

préparée à Laboratoire d'Informatique Fondamentale de Lille (LIFL)

soutenue le 18 Juillet 2011

Jury :

Rapporteurs :	Philippe Mulhem	CR1, CNRS (HDR)	Laboratoire d'Informatique de Grenoble
	Zhongfei Zhang	Professeur	Zhejiang University State University of New York
Directeur :	Chabane Djeraba	Professeur	Université Lille 1
Co-encadrant :	Jean Martinet	Maître de Conférences	Université Lille 1
Président :	Sophie Tison	Professeur	Université Lille 1
Examinateurs :	Bernard Merialdo	Professeur	Eurecom Sophia-Antipolis

T H E S I S

to obtain the title of

Doctor of Philosophy (PhD)

delivered by LILLE 1 University - Science and Technology

Specialty : Computer Science

Defended by

Ismail El Sayad

A Higher-level Visual Representation for Semantic Learning in Image Databases

prepared at Lille's Computer Science Laboratory (LIFL)

defended on July 18, 2011

Jury :

Reviewers :	Philippe Mulhem	CR1, CNRS (HDR)	Laboratoire d'Informatique de Grenoble
	Zhongfei Zhang	Professor	Zhejiang University
			State University of New York
Advisor :	Chabane Djeraba	Professor	Université Lille 1
Co-advisor :	Jean Martinet	Assistant Professor	Université Lille 1
President :	Sophie Tison	Professor	Université Lille 1
Examinator :	Bernard Merialdo	Professor	Eurecom Sophia-Antipolis

To my dear parents
To my grandparents
To my brothers
To my wife
To my son

I dedicate this thesis
Ismail

Acknowledgements

I would like to thank my supervisor Professor Chabane Djeraba and my co-supervisor Dr.Jean Martinet. Without their support, guidance and encouragement this work would not have been possible. Their commitment to hard work was a great source of inspiration over the last three years, and motivated me to always do my best. Their worthy advice and scrupulous assistance made me more conversant in the research field of image analysis and semantic understanding.
The tedious work of reviewing this dissertation was performed by Prof. Zhongfei (Mark) Zhang and Prof. Philippe Mulhem. I would like to give thanks for their valuable correction, suggestion, and acceptance. As well, thanks much the president, Prof. Sophie Tison, as well as the examiner, Prof. Bernard Merialdo, of my dissertation committee.
I would like to mention my long-term lab colleagues Samir, Taner, Thierry, Yassine, Marius, Rémi, Amal, Emilie, Céline, Tarek, Nacim for being helpful colleagues, friends and entertaining travel buddies. I would thank them for creating a friendly working environment and for the fruitful discussions and relaxing tea breaks. I am grateful to Emilie for proofreading parts of this thesis and translate them to French.
I would also like to thank all the new friends I have made at TU Delft during my internship in Netherland. I had with them interesting discussions that were helpful to improve my dissertation work.
Finally, I would like to thank my parents, wife, and brothers for their support and understanding. Specifically, I would acknowledge the support of my beloved wife who always being beside me in my dissertation work. Her love and support are greatest motivation for my PhD study. I owe so much to my lovely son Abdel kader who was born during my study and I did not spend so much time with him.

Abstract

With the availability of massive amounts of digital images in personal and on-line collections, effective techniques for navigating, indexing and searching images become more crucial. In this thesis, we rely on the image visual content as the main source of information to represent images. Starting from the bag of visual words (BOW) representation, a higher-level visual representation is learned where each image is modeled as a mixture of visual topics depicted in the image and related to high-level topics. First, we enhance the BOW representation by characterizing the spatial-color constitution of an image with a mixture of n Gaussians in the feature space. This leads to propose a novel descriptor, the *Edge Context*, which plays a role as a complementary descriptor in addition to the SURF descriptor. Such enhancements incorporate different image content information. Second, we introduce a new probabilistic topic model, *Multilayer Semantic Significance Analysis* (MSSA) model, in order to study a semantic inference of the constructed visual words. Consequently, we generate the *Semantically Significant Visual Words* (SSVWs). Third, we strengthen the discrimination power of SSVWs by constructing *Semantically Significant Visual Phrases* (SSVPs) from frequently co-occurring SSVWs that are semantically coherent. We partially bridge the intra-class visual diversity of the images by re-indexing the SSVWs and the SSVPs based on their distributional clustering. This leads to generate a *Semantically Significant Invariant Visual Glossary* (SSIVG) representation. Finally, we propose a new spatial weighting scheme and a Multiclass Vote-Based Classifier (MVBC) based on the proposed SSIVG representation. The large-scale extensive experimental results show that the proposed higher-level visual representation outperforms the traditional part-based image representations in retrieval, classification and object recognition.

Keyword: Image Representation, Image Indexing, Bag of Visual Words (BOW), Probabilistic Topic Model , Weighting Scheme, Image classification, Image Retrieval, Object Recognition.

Résumé

Avec l'augmentation exponentielle de nombre d'images disponibles sur Internet, le besoin en outils efficaces d'indexation et de recherche d'images est devenu important. Dans cette thèse, nous nous baserons sur le contenu visuel des images comme source principale d'informations pour leur représentation. Basés sur l'approche des sacs de mots visuels, nous proposons une représentation visuelle avancée. Chaque image est modélisée par un mélange de catégories visuelles sémantiques, reliées à des catégories de haut niveau. Dans un premier temps, nous améliorons l'approche des sacs de mots visuels en caractérisant la constitution spatio-colorimétrique d'une image par le biais d'un mélange de n Gaussiennes dans l'espace de caractéristiques. Cela permet de proposer un nouveau descripteur de contour qui joue un rôle complémentaire avec le descripteur SURF. Cette proposition nous permet de résoudre le problème lié à la perte d'informations spatiales des sacs de mots visuels, et d'incorporer différentes informations relatives au contenu de l'image. Dans un deuxième temps, nous introduisons un nouveau modèle probabiliste basé sur les catégories : le modèle MSSA (*Multilayer Semantic Significance Analysis* ou *Analyse multi-niveaux de la pertinence sémantique*) dans le but d'étudier la sémantique des mots visuels construits. Ce modèle permet de construire des mots visuels sémantiquement cohérents (SSVW - *Semantically Significant Visual Word*). Ensuite, nous renforçons la capacité de catégorisation des SSVW en construisant des phrases visuelles sémantiquement cohérentes (SSVP - *Semantically Significant Visual Phrase*), à partir des SSVW qui apparaissent fréquemment. Nous améliorons également l'invariance intra-classes des SSVW et des SSVP en les indexant en fonction de leur répartition, ce qui nous amène à générer une représentation d'un glossaire visuel invariant et sémantiquement cohérent (SSIVG -*Semantically Significant Invariant Visual Glossary*). Enfin, nous proposons un nouveau schéma de pondération spatiale ainsi qu'un classifieur multi-classes basé sur un vote. Nos résultats expérimentaux extensifs démontrent que la représentation visuelle proposée permet d'atteindre de meilleures performances comparativement aux représentations traditionnelles utilisées dans le domaine de la recherche, la classification et de la reconnaissance d'objets.

Mots-clés : Représentation d'images, Indexation d'images, Sacs de mots visuels, Modèle probabiliste, Pondération, Classification d'images, Reconnaissance d'objets.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Contents

1.1 Motivation

With the increasing convenience of capturing devices and the wide availability of large capacity storage devices, the amount of digital images that ordinary people can access has become vast. For example, it was reported that the FlickrTM photo repository has been hosting more than 5 billion images in September 2010 [1]. This huge amount is useless if there exists no effective technique for navigating, classifying and searching images.

The usual way to solve this problem consists in describing images by keywords,

1. http://edition.cnn.com/2010/TECH/web/09/20/flickr.5.billion.

and to use them in a *keyword-based* information retrieval or classification system. This method suffers from different drawbacks as follows.

First, it suffers from subjectivity and text ambiguity as they usually reflect the author's personal interpretation with respect to the image content. Figure 1.1 shows some examples of the retrieved images using the tag *jet* in Google image search engine [1]. If the user is searching for jet planes, then only a fraction of the images depicts the *jet* as might be expected. Instead, the tag sometimes denotes a jet engine or the Australian rock band. Thus, many photos are retrieved with no real common semantic theme.

Second, it requires a huge amount of time to manually annotate a whole database where no associated text is available for the images, as for instance many users do not annotate their pictures in their personal photo collection.

Nowadays images can be automatically described which only depends on their objective visual content [58]. When considering the visual contents of images, the problem of the semantic gap arises. The notion of semantic gap has been defined a decade ago by Smeulders et al. [141] as the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation. This lack of coincidence occurs due to the difference between the way we perceive visual content and what the machine can extract and define as shown in Figure 1.2. Hence, a semantic visual representation of images can help diminishing this gap. Moreover, this visual representation is applicable to effective techniques for navigating, classifying and searching images within a large-scale image database.

1. This query was performed at 12/05/2011 via Google image search engine website (http://images.google.com).

Figure 1.1: Examples of the Google image search results for *jet*.

1.2 Objectives

Recent research has shown that the part-based representation performance is much superior to the traditional image-based representation performance in the context of image retrieval and classification since a single global image feature is computed in the later which is not sufficient to represent the important local characteristics of objects [90]. Specifically, the bag of visual words (BOW) image representation [137] has drawn much attention, as it tends to code the local visual characteristics towards the object level, which is closer to the perception of the human visual systems [168].

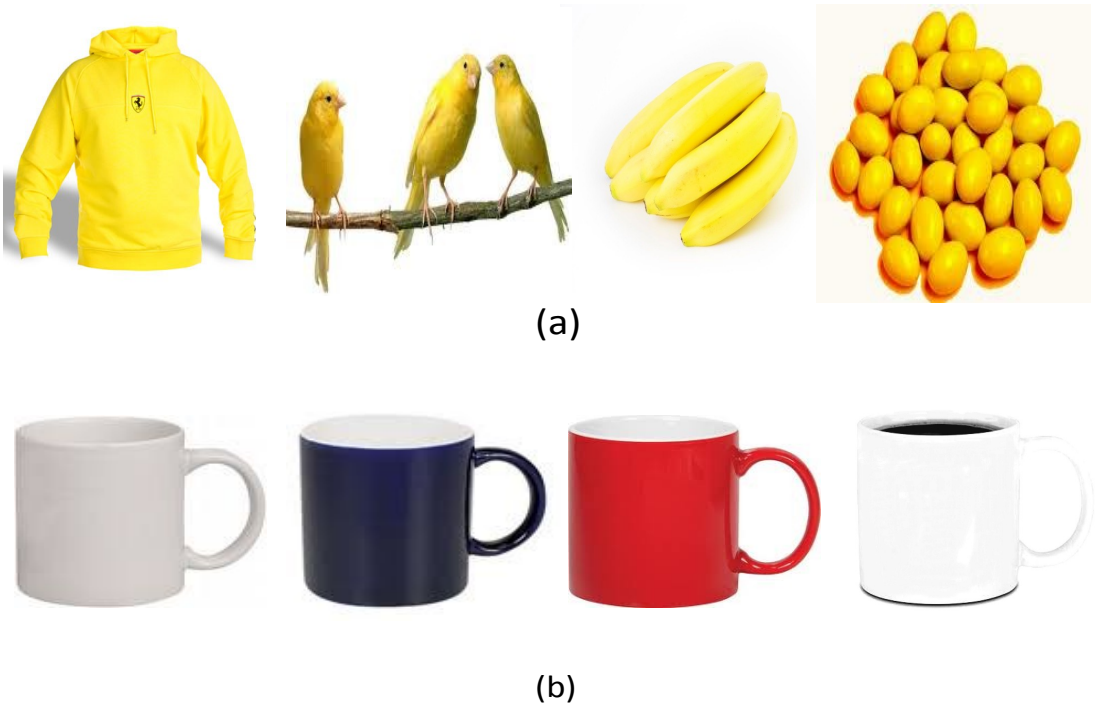

Figure 1.2: Illustration of the semantic gap: (a) Color-based features are unable to differentiate between these images representing different concepts. (b) Shape-based features cannot differentiate between the different cups.

Besides the good performance of the BOW representation, there are still drawbacks to be considered. The objectives of this work are to enhance the BOW representation and build a semantic higher-level visual representation that satisfies the following requirements:

1. The required visual representation needs to be based on local descriptors that describe the variety of local visual content information rather than describing only the intensity of the pixels. Most of the BOW representations use keypoints-based descriptors that make more direct use of pixel intensity values [89] such as SIFT [92], SURF [11], etc. This turns out to be pretty much of a challenge since these techniques do not handle well the large dis-

tortions that are due to pose and illumination variations. Furthermore, all other information like shapes and colors are disregarded with these descriptors, which is critical for many tasks such as handwritten digit recognition [82, 23] and face recognition [107].

2. The required visual representation is ought to be *less noisy* by eliminating the irrelevant or noisy visual words according to a semantic criterion that should also be defined. The noisy visual words are due to the visual vocabulary creation process in the BOW image representation. Such noisy visual words add ambiguity in the image representation.

3. The low discrimination power of the visual words leads to low correlations between the image features and the associated semantics. This is similar to the polysemy problem in a text, where one word can have different semantic meanings. Hence, the required higher-level visual representation needs to be *more discriminative* than the lower-level representation (BOW).

4. The required visual representation needs to be *invariant* to the visual diversity that is due to the arbitrary difference in visual appearances and shapes between the images of the same semantic class. Such visual diversity of objects causes one visual semantic to be represented by different visual representation units.

5. The *spatial constitution* of the images is desired to be included within the required visual representation since images are particular arrangements of patches in 2D space and the relative spatial relationship is an important factor in deriving semantic features [27].

6. The required visual representation should be relevant and useful to different

large-scale image retrieval and classification applications.

1.3 Contributions

Figure 1.3 illustrates the different hierarchal processes that enhance the BOW representation and generate the higher-level visual representation according to the different contributions highlighted as follows.

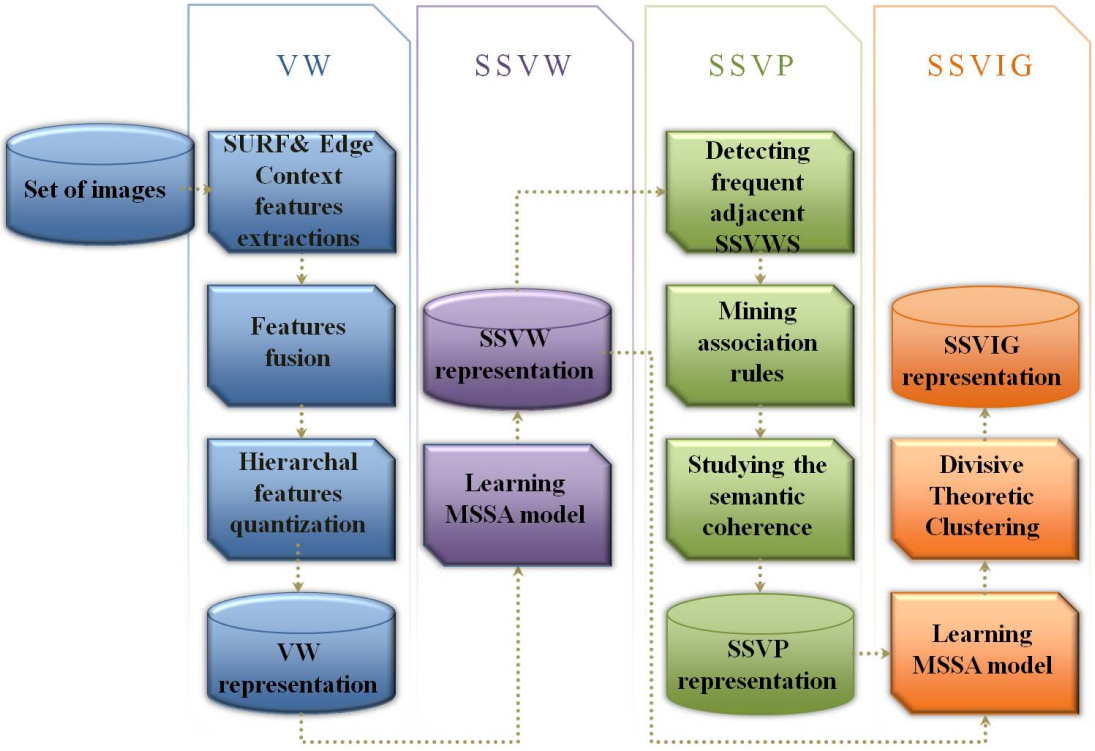

Figure 1.3: Different processes that generate the higher-level visual representation

Edge Context descriptor: We propose a novel descriptor, the *Edge Context*, that plays a role of complementary descriptor in addition to the SURF descriptor. We model the color-spatial constitution of an image with a mixture of n Gaussians

in a 5D color-spatial feature space. The Edge Context describes the distribution of the edge points at each detected interest point that are in the same Gaussian cluster by returning to the 5D color-spatial feature space.

Multilayer Semantic Significance Analysis (MSSA) model: Capturing the essential statistical characteristics of different visual representation units gives to the images a new representation, which is often more thrifty and less noisy. In our approach, we introduce a new probabilistic topic model, the *Multilayer Semantic Significance Analysis* (MSSA). This model differs from the pLSA model [59] and the LDA [16] model by introducing two layers of latent topics: high and visual latent topics. One layer represents the high aspects (i.e., image categories) and the other one represents the visual aspects (i.e., objects, parts of objects or scenes). We also make use of the MSSA in order to study the semantic inference of the different higher-level visual representation units.

Semantically Significant Visual Phrase (SSVW): The feature quantization process in BOW representation generates lots of unnecessary and insignificant visual words which are noisy in retrieval and classification. In our approach, Semantically Significant Visual Words (SSVWs) are selected from the constructed visual words based on their probability distributions to the relevant visual latent topics. Their probability distributions are estimated using the MSSA model.

Semantically Significant Visual Phrase (SSVP): In order to tackle the *low discrimination power* of the constructed SSVWs, we build a higher-level representation, named the *Semantically Significant Visual Phrase* (SSVP) from groups of adjacent significant visual words that frequently co-occur, being involved in strong

association rules [1], and semantically coherent.

Semantically Significant Invariant Visual Glossary (SSIVG): Studying the co-occurrence and the spatial scatter information makes the image representation more distinctive, the invariance power of the SSVWs and the SSVPs is still low. In our perception, that relevance-consistent group of the SSVWs or the SSVPs with similar semantic inferences should have the same index. Based on this, we cluster the SSVWs and the SSVPS using their probability distribution that are estimated using the MSSA model. After a distributional clustering, *each group of SSVWs that belongs to the same cluster are re-indexed with the same index as the cluster's centroid*. This step generates the Semantically Significant Invariant Visual Words (SSIVW), which consist of re-indexed SSVWs. In the same manner, we generate the Semantically Significant Invariant Visual Phrase (SSIVP). Finally, the SSIVWs and SSIVPs form the final visual representation, which is the Semantically Significant Invariant Visual Glossary (SSIVG) representation.

Novel spatial weighting scheme: An evident drawback of the BOW representation is the spatial information loss since the bag-of-visual-words approach represents an image as a collection of local patches ignoring their spatial structure within the image. To overcome this drawback, we propose a novel spatial weighting scheme for the SSIVW representation layer.

Multilayer Vote-Based Classifier (MVBC): Based on the proposed hierarchal representation, a new vote-based classifier, the MVBC, is introduced for classification and object recognition. This vote-based classifier is based on the

voting score of the SSIVWs and SSIVPs towards the dominant high topic in each test image.

Extensive experimental evaluations: We have conducted large-scale, extensive experimental performance evaluations of image retrieval, classification, and object recognition in comparison with various state-of-the-art image representation methods from the recent literature so as to demonstrate the superiority of the proposed higher-level visual representation methods.

1.4 Organization of the thesis

The structure of the thesis is as follows. Chapter 2 reviews the existing literature that focuses on different visual representations that have been proposed recently. Chapter 3 presents different probabilistic topic models and how they are applied to images. In this chapter, we also present the relation between Non-Negative Matrix Factorization (NMF) and topic models. Chapter 4 describes the vector space image model. It also reviews the different weighting schemes and the similarity measures that are used within the vector space model. Chapter 5 introduces the different techniques that we develop in order to enhance the BOW representation. In Chapter 6, we bring in the new probabilistic topic model, i.e., the MSSA model. In Chapter 7, we describe the different visual representation layers that lead to the SSIVG representation. The thesis is concluded in Chapter 8 with a brief statement, that presents the main contributions in a concise form, as well as some directions for further investigations.

Part I

Literature Review

Chapter 2

Visual Representation

Contents

2.1 Overview

Low-level image feature extraction is the basis of any visual representation. Low level features can be either extracted from the entire image or from local regions. Since there is no direct link between the high-level concepts and the low-level features [132], a semantic gap [140] appears and many visual representation

are proposed to bridge this gap.

This chapter is organized as follows. We briefly review low-level image features such as color, texture, shape and spatial location in Section 2.2. We discuss the notion of semantic gap of the semantic gap in image representation in Section 2.3. In order to bridge the semantic gap, many visual representations that are based on the low level features are introduced. These representations can be generally classified in two categories, image-based and part-based image representations. In Section 2.4, we review different image-based representations which are based on global feature descriptors over the whole image like color, color moment, shape or texture [44] global histograms. In Section 2.5, we review different part-based representations which are based on the statistics of features extracted from segmented image regions, salient key points or blobs [26, 72, 143, 145]. We give a summary and conclusion for this chapter in Section 2.6.

2.2 Low-level image features

Low-level feature extraction is a central pre-step for any visual representation. There are various kinds of features and each expresses a different aspect of a visual document [89]. This section gives a brief overview of existing feature classes, which are currently used such as color, texture, shape, or spatial location that can be extracted from the segmented regions.

2.2.1 Color feature

Color feature is one of the most widely used features in image representation. Colors are defined on a selected color space. Variety of color spaces are avail-

able, they often serve for different applications [65]. Color spaces shown to be closer to human perception and widely used in CBIR such as RGB, LAB, LUV, HSV (HSL), YCrCb and the hue-min-max-difference (HMMD) [98, 103, 91, 136]. Common color features or descriptors include, color-covariance matrix, color histogram, color moments, and color coherence vector [71, 156, 160, 176]. MPEG-7 has included dominant color, color structure, scalable color, and color layout as color features [126]. Gevers et al. [51] are interested in objects taken from different pointS of view and illumination. As the result, a set of viewpoint invariant color features have been computed. The color invariants are constructed on the basis of hue, hue-hue pair and three color features computed from reflection model.

Most of those color features though efficient in describing colors, are not directly related to high-level semantics. For a convenient mapping of region color to high-level semantic color names, some systems use the average color of all pixels in a region as its color feature [62, 103, 162]. Although most segmentation tends to provide homogeneous, color regions, due to the inaccuracy of segmentation, average color could be visually different from that of the original region. In Liu et al. [91], a dominant color in HSV space is defined as the perceptual color of a region. To obtain the dominant color, the authors first calculate the HSV space color histogram ($10 \times 4 \times 4$ bins) of a region and select the bin with maximum size. The average HSV value of all the pixels in the selected bin is defined as the dominant color. It is observed that in most cases, the average color and the dominant color are very similar, as in Figure 2.1(1). However, in some cases, they can be visually very different as in Figure 2.1(2).

The selection of color features depends on the segmentation results. For instance, if the segmentation provides regions, which do not have homogeneous

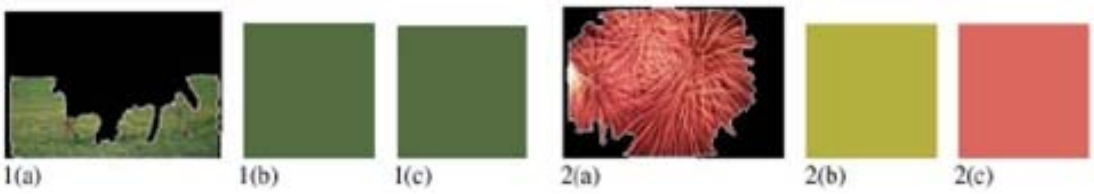

Figure 2.1: Average color and dominant color: (a) original region; (b) average color; (c) dominant color.

color, obviously the average color is not a good choice. It is stated that for more specific applications such as human face database, dominant knowledge can be explored to assign a weight to each pixel in computing the region colors [62].

It should be noted that in most of the CBIR works, the color images are not pre-processed. Since color images are often corrupted with noise due to capturing devices or sensors, it would improve retrieval accuracy significantly if an effective filter was applied to remove the color noise. A number of such color filters are available for this purpose [119, 118, 95].

2.2.2 Texture feature

Texture is not as well-defined as color features, some systems do not use texture features [147, 62, 103, 146]. However, texture provides important information in image classification as it help describing the content of many real-world images such as fruit skin, clouds, trees, bricks, or fabric. Hence, texture is an important feature in defining high-level semantics for image representation.

Texture features include spectral features, such as those obtained using Gabor filtering [96] or wavelet transform [161], statistical features characterizing texture in terms of local statistical measures, such as the six Tamura texture features

[155], and wold features proposed by Liu et al. [88]. Among the six Tamura features: coarseness, directionality, regularity, contrast, line-likeness, and roughness. The first three features are more significant [155]. The other three are related to the first three and do not add much to the effectiveness of texture description. MPEG-7 has employed the regularity, directionality and coarseness as the texture browsing descriptor [98, 126]. The wold features of periodicity, randomness and directionality have been proved to work well on Brodatz textures [17].

The limitation of Tamura features is that there has been no work at multiple resolutions to account for scale. Wold feature is also affected by image distortions such as scale and orientation variations due to perspective distortion [165]. Though working well on Brodatz textures, these features are proved to be less effective when applied to natural scene image retrieval and classification as texture regions in such images are less structured and homogeneous [165].

Among the various texture features, Gabor features and wavelet features are widely used for visual representation and have been reported to well match the results of human vision study [96, 161, 126]. Gabor filtering and wavelet transform are originally designed for rectangular images. However, natural images are of arbitrary-shapes. Hence, it is problematic to extract texture features from arbitrary-shaped images.

Texture features are usually obtained based on the texture property of pixels or small blocks contained in a local region of an image. For example, Ma et al.[96] use for each region, the mean value of the texture features of all the 4 $\times$ 4 blocks it contains as the region feature. The problem of such feature is that they cannot sufficiently describe the texture property of the entire region. An intuitive way to solve this problem is to extend the arbitrary-shaped region into

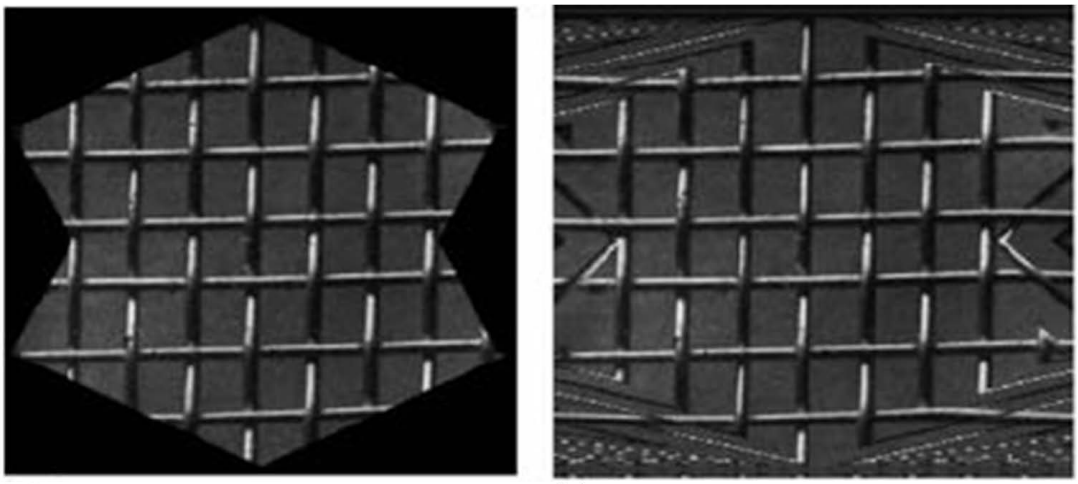

Figure 2.2: Arbitrary-shaped region and padded results: (a) original region; (b) mirroring padded result.

a rectangular area by padding some values outside the boundary and then apply block transforms. However, as regions in real-world images do not usually have homogeneous texture, such initial padding will introduce spurious components that do not describe the original region which degrades the accuracy quality of the texture feature obtained. Figure 2.2 gives an example of initial padding. Another texture descriptor based on local regions is the Edge Histogram Descriptor (EHD). It is found to be quite effective for representing natural images [126]. It captures the spatial distribution of edges, somewhat in the same idea as the color layout descriptor. To compute the EHD, a given image is first sub-divided into 4×4 sub-images, and local edge histograms for each of these sub-images is computed. Edges are broadly grouped into five categories: vertical, horizontal, 459°, 135° and neutral. Thus, each local histogram has five bins corresponding to the above five categories. The image partitioned into 16 sub-images resulting in 80 bins. These

bins are non-uniformly quantized using 3 bits/bin, resulting in a descriptor of size 240 bits. The EHD gives a rather precise description of the edge distribution But the EHD can be *very sensitive* to objects or scene distortions.

Zabih and Woodll [158] have developed a texture descriptor robust to illumination changes. It relies on histograms of ordering and reciprocal relations between pixel intensities which are more robust than raw pixel intensities. The binary relations between intensities of several neighboring pixels are encoded by binary strings and a distribution of all possible combinations is represented by histograms. This descriptor is suitable for texture representation but a *large number of dimensions* is required to build a reliable descriptor [114].

Lowe [93], proposed a scale invariant feature transform (SIFT), which combines a scale invariant interest point detector and a descriptor based on the gradient distribution in the neighboring regions.

First a set of orientation histograms are created on 4×4 pixel neighborhoods with 8 bins each. These histograms are computed from magnitude and orientation values of samples in a 16×16 region around the keypoint such that each histogram contains samples from a 4×4 sub-region of the original neighborhood region. The magnitudes are further weighted by a Gaussian function with s equal to one half the width of the descriptor window. The descriptor then becomes a vector of all the values of these histograms. Since there are $4 \times 4 = 16$ histograms each with 8 bins, the vector has 128 elements. This vector is then normalized to unit length in order to enhance invariance to affine changes in illumination. To reduce the effects of non-linear illumination a threshold of 0.2 is applied and the vector is again normalized.

The good performance of SIFT compared to other descriptors [106] is remark-

able. It is mixing of crudely localized information and the distribution of gradient related features seems to yield a good distinctive power while fending off the effects of localization errors in terms of scale or space. Using relative strengths and orientations of gradients reduces the effect of photometric changes.

More recently, Bay et al. [11] proposed a novel detector-descriptor scheme, SURF (Speeded-Up Robust Features), that is similar to SIFT, with a complexity stripped down even further in order to increase the efficiency. The detector is based on the Hessian matrix , but uses a very basic approximation, just as the DoG is a very basic Laplacian-based detector. It relies on integral images to reduce the computation time and they therefore call it the Fast-Hessian detector. The descriptor describes a distribution of Haar-wavelet responses within the interest point neighborhood. The integral images are exploited to speed up the process. Moreover, only 64 dimensions are used, reducing the time for feature computation and matching, and increasing simultaneously the robustness.

2.2.3 Shape feature

Shape is a fairly well-defined concept. Shape features of general applicability include aspect ratio, circularity, Fourier descriptors, moment invariants, consecutive boundary segments [102], etc. Shape features are important image features though they have not been widely used in RBIR as color and texture features. Shape features have shown to be useful in many domain specific images such as man-made objects. For color images used in most evaluations, however, it is difficult to apply shape features compared to color and texture due to the inaccuracy of segmentation. Despite the difficulty, shape features are used in some systems

and have shown potential benefit for RBIR. For example, Mezaris et al. [103] simple shape features such as eccentricity and orientation are used. The system introduced by Wang et al. [160] uses normalized inertia of order $1-3$ to describe region shape. Town at al. [156] introduce gross region shape descriptors based on area and second-order moments are used. MPEG-7 has included three shape descriptors for object-based image retrieval, the first one is the 3-D shape descriptor derived from 3-D meshes of shape surface, the second one is for region-based shape derived from Zernik moments, and the other is for contour based shape derived from curvature scale space (CSS) [126].

Although the CSS descriptor is invariant to translation, scaling and rotation, it is sensitive to general distortions, which can result from objects taken from different point of view. Mokhtarian and Abbasi have extended the CSS descriptor to be robust to affine transform which is a common way to approximate general shape distortions [110].

2.2.4 Spatial location

Besides color and texture, spatial location is also useful in image retrieval and classification. For example, 'sky' and 'sea' could have similar color and texture features, but their spatial locations are different with the sky usually appears at the top of an image, while sea lies at the bottom.

Spatial location are usually simply defined as *upper, bottom, top* according to the location of the local region in an image [144, 108]. In the approach introduced by Ma et al. [96], the region centroid and its minimum bounding rectangle are used to provide spatial location information. Mezaris et al. [103] use the spatial

center of a region to represent its spatial location.

Relative spatial relationship is more important than absolute spatial location in deriving semantic features. 2D-string [27] and its variants are the most common structure used to represent directional relationships between objects such as *left/right*, *below/above*. However, such directional relationships alone are not sufficient to represent the semantic content of images ignoring the topological relationships.

To better support semantic-based image retrieval, a spatial context-modeling algorithm [97] is presented, which considers six spatial relationships between region pairs: left, right, up, down, touch and front. An interesting method was proposed by Smith et al. [142]. The system uses a Composite Region Template (CRT) to define the spatial arrangement of regions and each semantic class is characterized by the CRTs obtained from a collection of sample images [142].

2.3 Semantic gap

As we mentioned before, many sophisticated algorithms have been designed to describe color, shape, and texture features, these algorithms cannot adequately model the image semantics and have many limitations when dealing with broad content image databases [109]. Extensive experiments on these systems show that low-level contents often fail to describe the high level semantic concepts in user's mind [179]. Therefore, the quality of the visual representation based on these features is still far from user's expectations.

More specifically, the discrepancy between the limited descriptive power of low-level image features and the richness of user semantics, is referred to as the

semantic gap [140]. This semantic gap is due to the following inherent problems.

One problem is that the extraction of complete semantics from image data is extremely hard, as it demands general object recognition and scene understanding. Despite encouraging recent progress in object detection and recognition, unconstrained broad image domain remains a challenge for computer vision. For instance, consumer photographs exhibit highly varied contents and imperfect image quality due to spontaneous and casual nature of image capturing. The objects in consumer images are usually ill-posed, occluded, and cluttered with poor lighting, focus, and exposure. There is usually large number of object classes in this type of polysemic images. Robust object segmentation for such noisy images is still an open problem.

The other problem causing the semantic gap is the complexity, ambiguity and subjectivity in user interpretation. Relevance feedback is regarded as a promising technique to solicit user's interpretation at post-query interaction. However, the correctness of user's feedback may not be statistically reflected due to the small sampling problem.

Therefore, an ideal visual representation should provide full support in bridging the semantic gap between numerical image features and the richness of human semantics [179, 140]. Since the step from the low-level numerical image features to the high-level human semantics is not straightforward, many visual representations are proposed using different techniques based on several chains of processes, in order to extract and refine the information incrementally. In the following sections, we review different kinds of these visual representations.

2.4 Image-based visual representation

In the image-based visual representations, each image is represented by a single global feature capturing information from the whole image. A great amount of information regarding the constituents of the image, such as individual regions or objects is lost in these representations. Once each image's feature is computed, the similarity can be measured between any pair of images using some distance metric e.g. L2 distance.

Swain and Ballard [152] were the first to use global histograms as image-based visual representation. They realized that the power to identify an object using color is much larger than that of a gray-valued image. As a histogram loses all information about the location of an object in the image, Ennesser and Medioni [43] project the histogram back into the image to locate it by searching for best matches. A histogram may be effective for retrieval as long as there is uniqueness in the color pattern held against the pattern in the rest of the entire data set. Swain and Ballard also argue that color histograms change slowly with change in viewpoint and scale and with occlusion.

One of the major problems of the global histograms is the lack of scalability. When very large data sets are at stake, plain histogram comparison will saturate the discrimination. For a 64-bin histogram, experiments show that, for reasonable conditions, the discriminating power among images is limited to $25,000$ images [151]. To keep up performance, in [117], a joint histogram is used, providing discrimination among $250,000$ images in their database, yielding 80 percent recall among the best 10 for two shots from the same scene using simple features. Other joint histograms add local texture or local shape [55], directed edges [67], and local

higher order structures [47].

Another alternative is to add a dimension representing the local distance. This is the correlogram [63], defined as a three dimensional histogram where the colors of any pair are along the first and second dimension and the spatial distance between them is along the third. The auto correlogram defining the distances between pixels of identical colors is found on the diagonal of the correlogram. A more general version is the geometric histogram [123], with the normal histogram, the correlogram, and several alternatives as specific cases. This also includes the histogram of the triangular pixel values, reported to outperform all of the above as it contains more information.

The main drawback of such kind of representations is being based on global histograms that have high sensitivity to scale, pose and lighting condition changes, clutter and occlusions. In addition, the global histograms cannot capture the local information of an image, which is so important for many tasks in image retrieval. For example, for High Resolution Computed Tomographic (HRCT) images of the lung, a disease such as emphy-sema manifests itself in the form of a low-attenuation region that is textured differently from the rest of the lung. Local features are needed for such situations because the number of pathology bearing pixels in an image is small relative to the rest of the pixels and any global signature would not be sufficiently impacted to serve as a useful attribute for image retrieval.

2.5 Part-based visual representation

Many part-based image representations are proposed recently such as visterms [69, 122, 120], blobs [26], and VLAD [68] which is vector representation of an image which aggregates descriptors based on a locality criterion in the feature space. An alternative approach is proposed by Morand et al. [111]. This approach introduced a scalable object-based indexing method for video content by objects without parsing them into their constituent elements. Morand et al. built a representation based on multi-scale histograms of wavelet coefficients of objects. In this case, the performance of the whole system is closely related to the accuracy of the object extraction process.

Recently, there is a trend of using image local patches for visual representations in the context of image retrieval and classification [81, 46, 70, 137, 178]. The salient image patches contain rich local information about an image. They are automatically extracted after detecting interest points using various detectors [105] and described by low level features [106]. The extracted low level features of all the local patches are then grouped into a large number of clusters. This leads to construct visual vocabulary where a visual word is defined as follows.

Definition 1 (Visual Word (VW))**.** *A visual word is a local segment in an image, defined by a reference point together with its neighborhood and an index generated from the feature quantization process.*

With its extracted low level features mapped into visual words, an image can be represented as a *Bag of Visual Words* (BOW), or specifically, as a vector containing the (weighted) count of each visual word in that image, which is used as feature vector in the classification task.

This BOW image representation is analogous to the bag-of-words representation of text documents in terms of both form and semantics, which makes techniques for text representation readily applicable to the visual representation. The BOW representation has drawn much attention recently, as it tends to code the local visual characteristics toward object level and achieves good results in representing variable object appearances caused by changes in pose, scale and translations [173, 72]. However, as discussed in the Introduction, the BOW represenation suffers from some drawbacks.

In BOW representation, the vocabulary creation process, based on clustering algorithms such as k-means, is quite rude and leads to many noisy words. Such words add ambiguity in the image representation. This problem has been addressed in the first video-Google paper by Sivic and Zisserman [137]. They used stop-lists that remove the most and least frequent words from the collection. Yang et al. [168] pointed out the ineffectiveness of this method and proposed several measures usually used in feature selection for machine learning or text retrieval.

Another evident drawback in BOW representation is the spatial information loss. To overcome this, Lazebnik et al. [81] extended the BOW representation to Spatial Pyramid Matching Kernel (SPM) by exploiting the spatial information of location regions. Recently, Yang et al. [169] tackled the two drawbacks (quantization rudeness and spatial information loss) and proposed an extension of SPM by replacing K-Means with sparse coding. In sparse coding and feature selection techniques, local features are dealt separately. The mutual dependence and interrelation among local features are ignored. However, recent work shows that the relationships among the local features are important for image repre-

sentation, such as the geometric relationship [166]. Gao et al. [49] introduced Laplacian sparse coding to enhance the sparse coding by constructing a Laplacian matrix, which can well characterize the similarity between local features. This representation, however, lacks to semantic learning that would better characterize the semantic relationships between the visual words.

To address the discrimination problem of visual words, Zheng and Gao [175] made an analogy between image retrieval and text retrieval, and have proposed a higher-level representation (***visual phrase***) based on the analysis of visual word occurrences to retrieve images containing desired objects. ***Visual phrases*** are defined as pairs of adjacent local image patches. The motivation of the visual phrase is to have a compact representation, which has more discrimination power than the lower level (visual words). later, Zhang et al. [174] enhance this approach by selecting descriptive visual phrases from the constructed visual phrases according to the frequencies of their constituent visual word pairs. In these two approaches, the higher-level (visual phrase) is defined as adjacent pairs of visual words which do not necessary guarantee a truly meaningful descriptive visual representation [171]. In addition, there are ambiguities in visual word lexicons. If the generation of the representation is a pure bottom-up process, the imperfectness in the visual words would never be reduced, and the quantization error would never be corrected without a pre-filtering step for the visual words done at a lower level. Yuan et al. [172] have proposed another higher-level lexicon, i.e. visual phrase lexicon, where a visual phrase is a spatially co-occurrent pattern of visual words. This higher-level lexicon is much less ambiguous than the lower-level one (visual words). The main contribution of this approach is to present a fast solution to the discovery of significant spatial co-occurrent patterns using frequent item set

mining. Zheng et al. [177] proposed a similar approach by constructing another high-level, delta visual phrase, and grouped delta visual phrases according to their similarity to visual synsets. Both approaches evaluated the significance of the visual phrases statistically. Zheng et al. [177] addressed the importance of the semantic factor but they measured the significance of a delta visual phrase based on its frequency as well as the frequencies of its constituent visual words.

Hoíng et al. in [60] have proposed to construct another higher level representation (triplets of entities) from visual words (entities) by studying the spatial relationships between them. The proposed representation describes triangular spatial relationships with the aim of being invariant to image translation, rotation, scale, flipping, and robust to view point changes if required. Beside we share the same motivation for constructing a higher level representation, this approach lacks statistical and semantic learning for the lower level which is a pre-step to construct the higher level representation in our approach.

Our framework differs from these approaches by proposing the MSSA model to analyze the semantic significance of the visual words in order to overcome the rudeness of quantization. We also utilize MSSA to check the semantic coherence of groups of Semantically Significant Visual Words (SSVWs) that are spatially adjacent and frequently occur with each other in order to construct another higher-level representation named Semantically Significant Visual Phrase (SSVP). This representation is more discriminative and descriptive than SSVW.

2.6 Summary and conclusion

In this chapter, we briefly review different low-level image features such as color, texture, shape and spatial location. We discuss the definition of the semantic gap in visual representation. We review different visual representations are introduced in order to bridge the semantic gap. These visual representations can be generally classified in two groups as follows.

- Image-based representations which are based on global feature descriptor over the whole image like color, color moment, shape or texture histograms [44].
- Part-based representations which are based on the statistics of features extracted from segmented image regions, key points or blobs [26, 72, 143, 145].

On the one hand, the global representations fail to capture the local information of an image, which is so important for many tasks in image retrieval and classification. On the other hand, the part-image representation especially the bag of visual words (BOW) image representation [137] has drawn much attention, as it tends to code the local visual characteristics towards the object level, which is closer to the perception of human visual systems [174]. Beside, the significant performance of the BOW representation, still there are drawbacks to be considered such spatial information loss, feature quantization nosiness, low discrimination power.

Chapter 3

Probabilistic Topic Models for Semantic Learning

Contents

3.1 Overview

After we have reviewed different kinds of the visual representations in the previous chapter, we review in this chapter the key ideas behind the topic models and how they capture the essential statistical characteristics of the visual representation units. By capturing and learning the statistical characteristics of the visual representation units, then one gives the images a new representation, which is often more parsimonious and less noise-sensitive. Probabilistic topic models extract a set of latent topics from a corpus and as a consequence represent the images in a new latent semantic space. One of the well-known topic models is the Probabilistic Latent Semantic Analysis (pLSA) model proposed by Hofmann [59] for text document semantic analysis and is applied later to images. In pLSA each image is modeled as a probabilistic mixture of a set of topics. Going beyond PLSA, Blei et al. [16] presented the Latent Dirichlet Allocation (LDA) model by incorporating a prior for the topic distributions. In these probabilistic topic models, one assumption underpinning the generative process is that images are independent and one layer of topics are proposed.

This chapter is organized as follows. In Section 3.2, we review the probabilistic generative process that describes how words in documents might be generated on the basis of latent variables. In Section 3.3, we review the different topic models that is proposed for text document and we describe how they are applied in images. We discuss the graphical notations that are used to represent the different probabilistic topic models in Section 3.4. We describe the geometrical interpretation of the probabilistic topic models in Section 3.5. In Section 3.6, we review the relation between probabilistic topic models and Non-negative Matrix

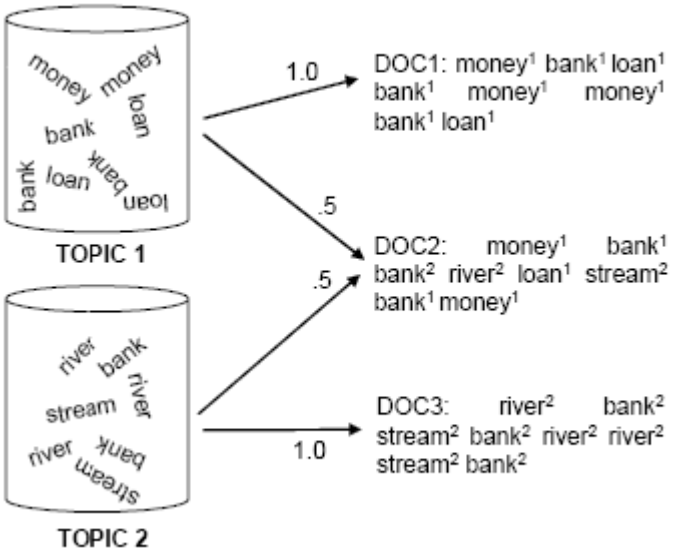

Figure 3.1: Illustration of the generative process (from [148]).

Factorization (NMF). We summarize this chapter and give a conclusion in Section 3.7.

3.2 Probabilistic generative process

By returning to text documents where words constitute the elementary parts of the documents, a generative model for documents is based on simple probabilistic sampling rules that describe how words in documents might be generated on the basis of latent (random) variables. When fitting a generative model, the goal is to find the best set of latent variables that can explain the observed data (i.e., observed words in text documents), assuming that the model actually generated the data.

Figure 3.1 illustrates the probabilistic generative process with two topics and

three text documents. Topics 1 and 2 are thematically related to money and rivers and are illustrated as bags containing different distributions over words. Different documents can be produced by picking words from a topic depending on the weight given to the topic. For example, documents 1 and 3 were generated by sampling only from topic 1 and 2 respectively while document 2 was generated by an equal mixture of the two topics. Note that the superscript numbers associated with the words in documents indicate which topic was used to sample the word. The way that the model is defined, there is no notion of mutual exclusivity that restricts words to be part of one topic only. This allows topic models to capture polysemy, where the same word has multiple meanings. For example, both the money and river topic can give high probability to the word BANK, which is sensible given the polysemous nature of the word.

The generative process described here does not make any assumptions about the order of words as they appear in documents. The only information relevant to the model is the number of times words are produced. This is known as the bag-of-words assumption as we discussed before, and is common to many statistical language models including LSA [34].

This generative process is applied to images in the same manner and the visual words are used instead of the textual words. Of course, the spatial arrangements of the visual words are important cues to the content of an image and this information is not utilized by such generative process.

3.3 Different probabilistic topic models

By returning to the semantic learning in text documents, a variety of probabilistic topic models have been used to analyze the content of documents and the meaning of words [16, 53, 59]. These models all use the same fundamental idea that a document is a mixture of topics but make slightly different statistical assumptions. Each word w_i in a document (where the index refers to the i^{th} word token) is generated by first sampling a topic from the topic distribution, then choosing a word from the topic-word distribution. In this section we will review the topic models that are recently used for semantic learning in text documents and we describe how they are applied to images.

3.3.1 Probabilistic Latent Semantic Analysis (pLSA)

Hofmann [59] introduced the probabilistic topic approach to document modeling in his Probabilistic Latent Semantic Analysis (pLSA) method. The key concept of the pLSA model is to map the high dimensional word distribution vector of a document to a lower dimensional topic vector (also called aspect vector). It is assumed that each document consists of a mixture of multiple topics and that the occurrences of words are a result of the topic mixture. This generative model is expressed by the following probabilistic model:

$$P(d_j, w_i) = P(d_j)P(w_i|d_j) \tag{3.1}$$

$$P(w_i|d_j) = P(d_j)\sum_k P(z_k|d_j)P(w_i|z_k) \tag{3.2}$$

where $P(d_j)$ denotes the probability of a document d_j of the database to be picked, $P(z_k|d_j)$ the probability of a topic z_k given the current document, and $P(w_i|z_k)$ the probability of a word w_i given a topic. To simplify notations, let $\Phi^k = P(w|z_k)$ refer to the multinomial distribution over words for topic z_k and $\theta^j = P(z|d_j)$ refer to the multinomial distribution over topics for document d_j. The parameters Φ and θ indicate which words are important for which topic and which topics are important for a particular document, respectively

Once a topic mixture $P(z_k|d_j)$ is derived for each document d_j, a high-level representation based on the respective mode the words belong to has been found. At the same time this representation is of low dimensionality as commonly the number of concepts in the model is chosen to be much smaller than the number of words. The K-dimensional topic vector can be used directly as an index in an IR.

3.3.2 Latent Dirichlet Allocation (LDA)

Beside the good results of pLSA in document retrieval, the pLSA model does not make any assumption about how the mixture weights θ are generated. This makes pLSA fail to generate new documents which are not available in the training stage. Blei et al. [16] extended this model by introducing a Dirichlet prior on θ, calling the resulting generative model Latent Dirichlet Allocation (LDA). As a conjugate prior for the multinomial, the Dirichlet distribution is a convenient choice as prior, simplifying the problem of statistical inference. The probability density of a K dimensional Dirichlet distribution over the multinomial distribution $p = (p_1, ..., p_K)$ is defined by:

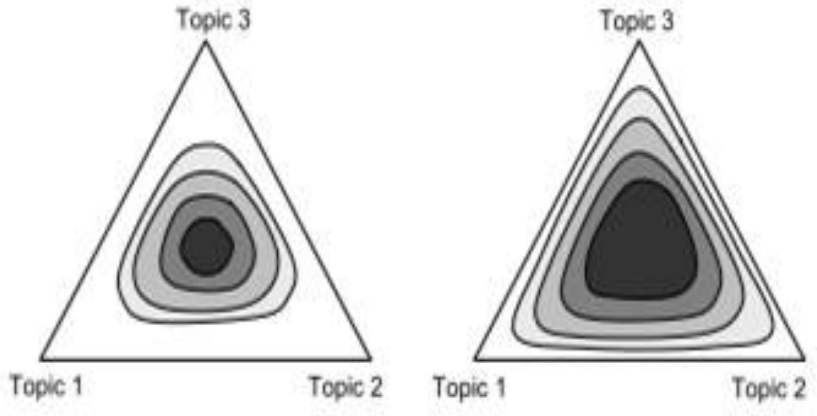

Figure 3.2: Illustration for the symmetric Dirichlet distribution for three topics on a two-dimensional simplex. Darker colors indicate higher probability. Left: α = 4, right: $\alpha = 2$ [148].

$$Dir(\alpha_1, ..., \alpha_K) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{K} p_k^{\alpha_k - 1} \tag{3.3}$$

The parameters of this distribution are specified by $\alpha_1 ... \alpha_K$. Each hyper parameter α_k can be interpreted as a prior observation count for the number of times topic z_k is sampled in a document, before having observed any actual words from that document. It is convenient to use a symmetric Dirichlet distribution with a single hyper parameterα such that $\alpha_1 = \alpha_2 = ... = \alpha_K = \alpha$. Figure 3.2 illustrates the Dirichlet distribution for three topics in a two-dimensional simplex. By placing a Dirichlet prior on the topics distributions θ, the result is a smoothed topic distribution, with the amount of smoothing determined by the α parameter.

Griffiths and Steyvers [53] explored a variant of this model, discussed by Blei et al. [16] , by placing a symmetric Dirichlet(β) prior on Φ as well. The hyper parameter β can be interpreted as the prior observation count on the number

of times words are sampled from a topic before any word from the corpus is observed. This smoothes the word distribution in every topic, with the amount of smoothing determined by β. Good choices for the hyper parameters α and β will depend on the number of topics and vocabulary size.

3.3.3 Extended probabilistic topic models

The statistical model underlying the topic modeling approach has been extended to include other sources of information about documents. For example, Cohn and Hofmann [31] extended the pLSA model by integrating content and link information. In their model, the topics are associated not only with a probability distribution over terms, but also over hyperlinks or citations between documents.

Recently, Steyvers et al. [149] proposed the author-topic model, an extension of the LDA model that integrates authorship information with content. Instead of associating each document with a distribution over topics, the author-topic model associates each author with a distribution over topics and assumes each multi-authored document expresses a mixture of the authors topic mixtures.

3.3.4 Applying probabilistic topic models to images

When topic model are applied to images, each image represents a single visual document. These models have been applied directly to image tags, as image tags consist of words or for visual words. Many of these models introduce *only one* latent, i.e. unobservable, topic layer between the documents (i.e., images here) and the words. Even Cohn and Hofmann [31] and Steyvers et al. [149], who link another type of information (citation and authors information) to the content of

the documents do not construct another topic layer.

They simply treat the linked information in the same way as the content not as new hidden topics. However, every image consists of one or more visual aspects (multiple objects parts or multiple objects), which in turn are combined to one or more higher-level aspects (i.e., visual category).

Lienhart et al. [86] introduced a new model named multilayer multimodal probabilistic Latent Semantic Analysis (mm-pLSA). They derive the training and inference rules for the smallest possible non-degenerated mm-pLSA model: a model with two leaf-pLSAs (here from two different data modalities: image tags and visual image features) and a single top-level pLSA node merging the two leaf-pLSAs. From this derivation, it is obvious how to extend the learning and inference rules to more modalities and more layers. Even though this approach introduced a new multilayer inference rules, it uses an EM algorithm to derive the different parameters, which costs a high computational power for parameters initialization and estimation. In addition, this approach did not introduce any criterion to estimate the number of different latent variables.

3.4 Graphical notation

Probabilistic generative models with repeated sampling steps can be conveniently illustrated using plate notations [20]. In these graphical notations, shaded and unshaded variables indicate observed and latent (i.e., unobserved) variables respectively. The variables Φ and θ, as well as z (the assignment of word tokens to topics) are the three sets of variables that we would like to infer. The hyper parameters α and β are treated as constants in the model.

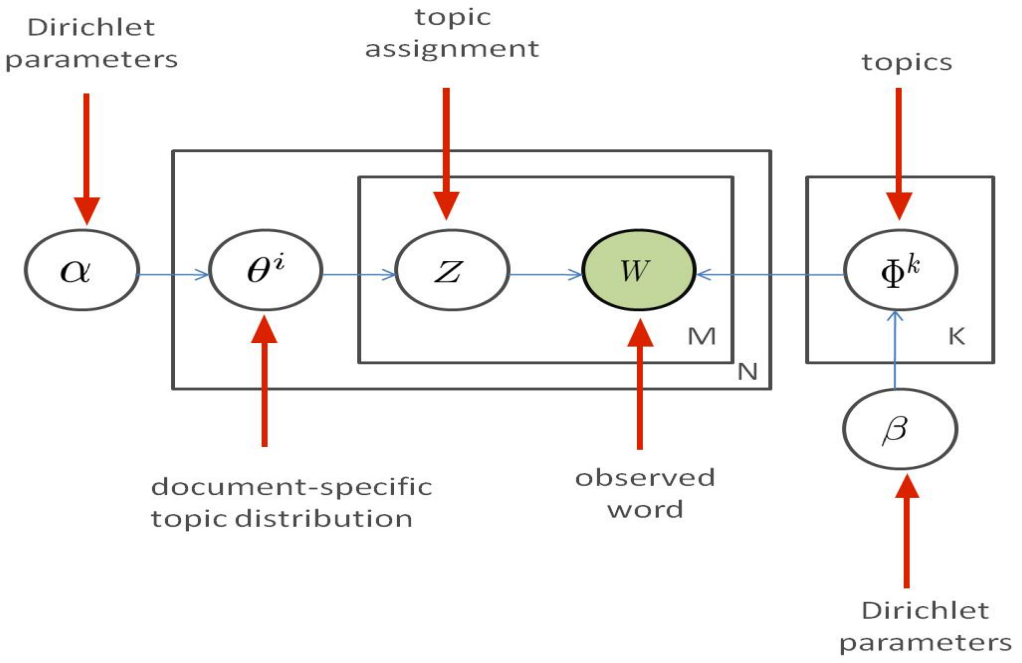

Figure 3.3: The graphical notation of a topic model .

Figure 3.3 shows the graphical notation of the topic model used in Griffiths and Steyvers [53]. Arrows indicate conditional dependencies between variables while plates (the boxes in the figure) refer to repetitions of sampling steps with the variable in the lower right corner referring to the number of samples. For example, the inner plate over z and w illustrates the repeated sampling of topics and words until M words have been generated for document d_j. The plate including θ^i illustrates the sampling of a distribution over topics for each document d for a total of N documents. The plate surrounding Φ^k illustrates the repeated sampling of word distributions for each topic z until K topics have been generated.

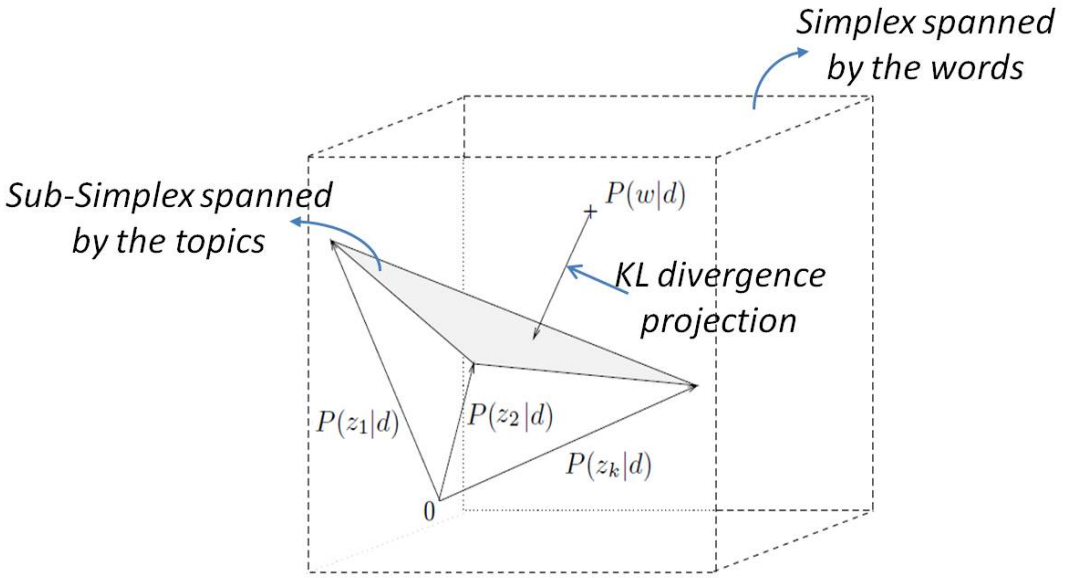

Figure 3.4: A geometric interpretation of the topic model (from Hofmann [59]).

3.5 Geometric interpretation

The probabilistic topic model has an elegant geometric interpretation as shown in Figure 3.4 that describes the geometric interpretation of pLSA model. With a vocabulary containing M words, an M dimensional space can be constructed where each axis represents the probability of observing a particular word type. The $M-1$ dimensional simplex represents all probability distributions over words.

Each document that is generated by the model is a convex combination of the K topics which not only places all word distributions generated by the model as points on the $K-1$ dimensional simplex, but also as points on the $K-1$ dimensional simplex spanned by the topics. The Dirichlet prior on the topic-word distributions can be interpreted as forces on the topic locations with a higher β moving the topic locations away from the corners of the simplex. When the

number of topics is much smaller than the number of word types (i.e., $K \ll M$), the topics span a low-dimensional subsimplex and the projection of each document onto the low-dimensional subsimplex can be thought of as dimensionality reduction. This formulation of the model is similar to Latent Semantic Analysis. Buntine [21] has pointed out formal correspondences between topic models and principal component analysis, a technique closely related to LSA.

3.6 Probabilistic topic models as Non-negative Matrix Factorization (NMF)

Non-negative Matrix Factorization (NMF) is one of the widely-used multivariate data analysis methods [115, 85, 83, 84, 130], which has many potential applications in pattern recognition and machine learning. NMF has been investigated by many researchers, e.g. Paatero and Tapper [115], but it has gained popularity through the work of Lee and Seung [83, 84]. Based on the argument that the non-negativity is important in human perception they proposed simple algorithms (often called the Lee-Seung algorithms) for finding non-negative representations of non-negative data and images.

In this section we describe the basic NMF model and review the multiplicative update rules that have been proposed recently for factorization. Finally, we describe the relation between the NMF and probabilistic topic models.

3.6.1 Basic NMF model

One common ground in the various approaches for noise removal, model reduction, feasibility reconstruction, image and text analysis and so on, is to replace the original data by a lower dimensional representation obtained via a subspace approximation. The use of low-rank approximations, therefore, comes to the forefront in a wide range of important applications. Factor Analysis and Principal Component Analysis are two of the many classical methods used to accomplish the goal of reducing the number of variables and detecting structures among the variables [14].

Often the data to be analyzed is non-negative, and the low-rank data are further required to be comprised of non-negative values in order to avoid contradicting physical realities. Classical tools cannot guarantee to maintain the non-negativity. The approach of finding reduced rank non-negative factors to approximate a given non-negative data matrix thus becomes a natural choice. This is the so-called Non-negative Matrix Factorization (NMF) problem which can be stated in a generic form as follows:

Definition 2. *(Non-negative Matrix Factorization (NMF))*

Given $A \in \mathbb{R}_+^{M\times N}$ *and a positive integer* $K \leq min\ (M\ ,\ N)$*, find* $W \in \mathbb{R}_+^{M\times K}$ *and* $H \in \mathbb{R}_+^{K\times N}$ *such that a divergence function* $D\ (A \parallel \tilde{A})$ *is minimized, where* $\tilde{A} = WH$ *is the reconstructed matrix from the factorization.*

NMF has been successfully applied to a variety of applications, including image part based representation [153], document clustering [167, 133], sound classification [28], medical imaging [85, 3], audio processing [76], bioinformatics [19], etc.

One of the prominent applications of NMF, which is of our interest, is the dyadic data analysis [133, 167]. Dyadic data refers to a domain with two finite sets of objects in which observations are made for dyads, i.e., pairs with one element from either set. In the simplest case of dyadic data - on which we focus - is an elementary observation consists just of (x_i, y_k) itself without any scalar value (strength of preference or association), i.e. a co-occurrence of $x_i \in X$ and $y_k \in Y$. In image part-based representation, X may correspond to a image collection, Y to visual terms vocabulary, and (xi, y_k) would represent the occurrence of a visual term y_k in a image x_i

3.6.2 Multiplicative update rules for factorization

Various divergence measures were considered as objective functions for factorization with non-negativity constraints, including sparseness constraints [61], Csiszàr's divergence [30], Bregman divergence [37], and a generalized divergence measure [78]. Two divergence measures that were considered by Lee and Seung [83, 84] are widely-used and are summarized as:

$$\varepsilon_1 = \|A - WH\|^2 = \sum_{m,n} [A_{mn} - [WH]_{mn}]^2 \tag{3.4}$$

$$\varepsilon_2 = \sum_{m,n} \left[A_{mn} log \frac{A_{mn}}{[WH]_{mn}} - A_{mn} + [WH]_{mn} \right]^2 \tag{3.5}$$

The minimization of the divergence functions described above, should be done with non-negativity constraints for both A and S. Multiplicative updating is an efficient way in such a case, since it can easily preserve non-negativity constraints

at each iteration. Multiplicative updating algorithms for NMF associated with these two objective functions are given as follows:

1. (LS) A local minimum of the objective function (3.4) is computed by the LS multiplicative algorithm that has the form

$$w_{mk} \leftarrow w_{mk} \frac{(AH^T)_{mk}}{(WHH^T)_{mk}} \tag{3.6}$$

$$H_{kn} \leftarrow H_{kn} \frac{(W^T A)_{kn}}{(W^T WH)_{kn}} \tag{3.7}$$

2. (I-divergence) In the case of I-divergence-based objective function (3.5), its minimum is found by the multiplicative updating algorithm that is of the form

$$W_{mk} \leftarrow W_{mk} \sum_n \frac{A_{mn}}{(WH)_{mn}} H_{kn} \tag{3.8}$$

$$H_{kn} \leftarrow H_{kn} \sum_m W_{mk} \frac{A_{mn}}{(WH)_{mn}} \tag{3.9}$$

Lee and Seung have shown that the application of the Multiplicative update rules in (3.6, 3.7) and (3.8, 3.9) are guaranteed to find at least locally optimal solutions of the objective functions (3.4) and (3.5), respectively. They have proven the convergence relying upon defining an appropriate auxiliary function. The multiplicative update rules themselves are extremely easy to implement;

3.6.3 Relation between probabilistic topic models and NMF

As we have mentioned that NMF and Probabilistic topic models have been successfully applied to a number of data analysis tasks. Despite their different inspirations, both methods are instances of multinomial PCA [22]. Gaussier and Goutte [50] have explored this relationship and first show that PLSA solves the problem of NMF with KL divergence, and then explore the implications of this relationship. Recently, Shashanka et al. [135] have shown that there are strong ties between non-negative matrix factorization and other probabilistic topic models, and provide some straightforward extensions which can help in dealing with shift invariances, higher-order decompositions and sparsity constraints.

The two dimensions of the matrix A can be presented by x_1 and x_2, respectively. The non-negative entries $A_{x_1 x_2}$ can be considered as having been generated by an underlying probability distribution $P(x_1, x_2)$. Variables x_1 and x_2 are multinomial random variables, where x_1 can take one out of a set of M values in a given draw and x_2 can take one out of a set of N values in a given draw. In other words, one can model A_{mn}, the entry in row m and column n, as the number of times features $x_1 = m$ and $x_2 = n$ were picked in a set of repeated draws from the distribution $P(x_1, x_2)$. Unlike NMF which tries to characterize the observed data directly, probabilistic topic models characterize the underlying distribution $P(x_1, x_2)$. There are two ways of modeling $P(x_1, x_2)$: symmetric and asymmetric factorization.

3.6.3.1 Symmetric factorization

Probabilistic topic models enable one to attribute the observations as being due to hidden or latent factors. The main characteristic of these models is conditional independence multivariate data are modeled as belonging to latent classes such that the random variables within a latent class are independently of one another. The model expresses a multivariate distribution such as $P(x_1, x_2)$ as a mixture where each component of the mixture is a product of one-dimensional marginal distributions. In the case of two dimensional data such as A, the model can be written mathematically as:

$$P(x_1, x_2) = \sum_{z \in \{1,2,...,K\}} P(z)P(x_1|z)P(x_2|z) \tag{3.10}$$

In (3.10), z is a latent variable that indexes the hidden components and takes values from the set $1, ..., K$. This equation assumes the principle of local independence, whereby the latent variable z renders the observed variables x_1 and x_2 independent. This model is presented independently as Probabilistic Latent Component Analysis ($PLCA$) [163]. The aim of the model is to characterize the distribution underlying the data as shown above by learning the parameters so that the hidden structure present in the data becomes explicit.

The model can be expressed as a matrix factorization. Representing the parameters $P(x_1|z)$, $P(x_2|z)$, and $P(z)$ as entries of matrices W, G, and S, respectively, where:

- W is an $M \times K$ matrix such that W_{mk} corresponds to the probability $\mathrm{P}(x_1 = m|z = k)$.
- G is a $K \times N$ matrix such that G_{kn} corresponds to the probability $P(x_2 =$

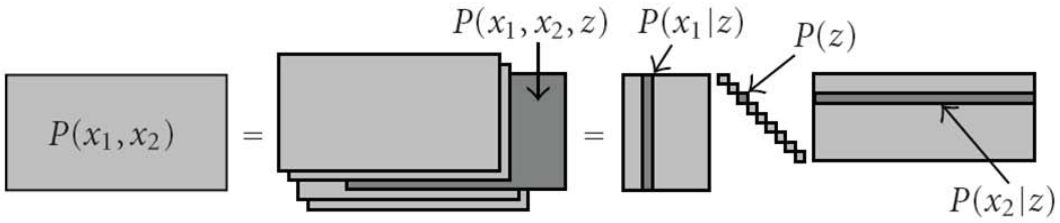

Figure 3.5: Probabilistic topic model of (3.10) as NMF (from Shashanka et al. [135]).

$n|z = k)$.

– S is a $K \times K$ diagonal matrix such that S_{kk} corresponds to the probability $P(z = k)$.

One can write the model of (3.10) in matrix form as:

$$P = WSG = WH \tag{3.11}$$

Where the entries of matrix P correspond to $P(x_1, x_2)$ and $H = SG$. Figure 3.5 illustrates the model schematically.

Parameters can be estimated using EM algorithm. The update equations for the parameters can be written as:

$$P(z|x_1, x_2) = \frac{P(z)P(x_1|z)P(x_2|z)}{\sum_z P(z)P(x_1|z)P(x_2|z)} \tag{3.12}$$

$$P(z|x_1, x_2) = \frac{\sum_{j\in\{1,2\}j\neq i,} A_{x_1x_2}P(z|x_1, x_2)}{\sum_{z,x_1,x_2} A_{x_1x_2}P(z|x_1, x_2)} \tag{3.13}$$

$$P(z) = \frac{\sum_{x_1,x_2,} A_{x_1x_2}P(z|x_1, x_2)}{\sum_{z,x_1,x_2} A_{x_1x_2}P(z|x_1, x_2)} \tag{3.14}$$

Writing the above update equations in matrix form using W and H from (3.11),

we obtain:

$$W_{mk} \leftarrow W_{mk} \sum_n \frac{A_{mn}}{(WH)_{mn}} H_{kn} \tag{3.15}$$

$$H_{kn} \leftarrow H_{kn} \sum_m W_{mk} \frac{A_{mn}}{(WH)_{mn}} \tag{3.16}$$

The above equations are identical to the NMF multiplicative update equations of (3.8) and (3.9) up to a scaling factor in H. This is due to the fact that the probabilistic model decomposes P which is equivalent to a normalized version of the data A. Smaragdis and Raj [163] present a detailed derivation of the update algorithms and a comparison with NMF update equations. This model has been used in analyzing image and audio data among other applications.

3.6.3.2 Asymmetric factorization

The probabilistic topic model of 3.10 considers each dimension symmetrically for factorization. The two dimensional distribution $P(x_1, x_2)$ is expressed as a mixture of two dimensional latent factors where each factor is a product of one-dimensional marginal distributions. Now, consider the following factorization of $P(x_1, x_2)$:

$$P(x_1, x_2) = P(x_i)P(x_j|x_i) \tag{3.17}$$

$$P(x_j|x_i) = \sum_z P(x_j|z)P(z|x_i) \tag{3.18}$$

Where i, $j \in 1, 2$, $i \neq j$ and z is a latent variable. This version of the model with asymmetric factorization is the same as PLSA model as discussed in Section 3.3.1.

Without loss of generality, let $j = 1$ and $i = 2$. We can write the above model in matrix form as $q_n = Wg_n$, where q_n is a column vector indicating $P(x_1|x_2)$, g_n is

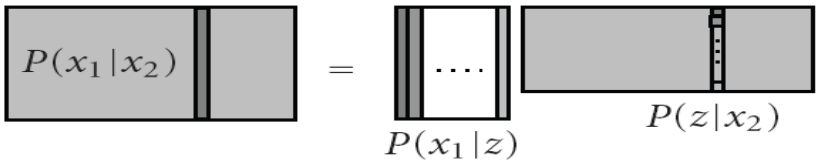

Figure 3.6: Probabllistic topic model of (3.17) as NMF (from Shashanka et al. [135]).

a column vector indicating $P(z|x_2)$, and W is a matrix with the (m, k)th element corresponding to $P(x1 = m|z = k)$. If z takes K values, W is an $M \times K$ matrix. Concatenating all column vectors q_n and g_n as matrices Q and G, respectively, one can write the model as:

$$Q = WG \tag{3.19}$$

$$A = WGS = WH \tag{3.20}$$

Where S is an $N \times N$ diagonal matrix whose n^th diagonal element is the sum of the entries of v_n (the n^th column of A), and $H = GS$. Figure 3.6 provides a schematic illustration of the model.

Given data matrix A, parameters $P(x_1|z)$ and P$(z|x_2)$ are estimated by iterations of equations derived using the EM algorithm:

$$P(z|x_1, x_2) = \frac{P(z|x_2)P(x_1|z)}{\sum_z P(z|x_2)P(x_1|z)} \tag{3.21}$$

$$P(x_1|z) = \frac{\sum_{x_2}, A_{x_1x_2}P(z|x_1, x_2)}{\sum_{z,x_1,x_2} A_{x_1x_2}P(z|x_1, x_2)} \tag{3.22}$$

$$P(z|x_2) = \frac{\sum_{x_1}, A_{x_1x_2}P(z|x_1, x_2)}{\sum_{x_1} A_{x_1x_2}} \tag{3.23}$$

Writing the above equations in matrix form using W and H from (3.20), we

obtain a set of Multiplicative update rules that are exactly to 3.12, 3.13 which are identical to the NMF multiplicative update equations of (3.8) and (3.9).

3.7 Summary and conclusion

Probabilistic topic models have the potential to make important contributions to the statistical analysis of large image collections. These models enable the use of sophisticated statistical methods to identify the underlying latent structure from a set of visual words. Consequently, it is easy to explore different representations of images, and to develop richer models capable of capturing more of the content. Topic models illustrate how using a different representation can provide new insights into the statistical modeling, incorporating many of the key assumptions behind LSA, and making it possible to identify a set of interpretable probabilistic topics rather than a semantic space.

All these models simply treat the images independently and one kind or layer of latent topics is proposed. However, every image consists of one or more visual aspects (multiple objects parts or multiple objects), which in turn are combined into one or more higher-level aspects (i.e., image category). For a test image, both of these aspects are hidden topics that are needed for image semantic analysis. We introduce a Multilayer Probabilistic Semantic (MSSA) model that consists of two layered of topics in order to select semantically significant visual words (SSVWs) from the classical visual words based on their probabilistic distributions to the relevant visual latent topics. In addition, we have utilized NMF to implement the proposed MSSA model and we have proposed new multiplicative update rules for factorization

Chapter 4

Image Indexing, Term Weighting and Similarity Measures

Contents

4.1 Overview

One of the successful aspects in text information retrieval has been the definition of the Vector Space Model [128] that may be further enriched with weighting schemes, static ordering, probabilistic models, and latent semantic indexing [25]. As we mentioned before, the analogy between text documents and images considers that an image is represented as a bag of visual words. Following this analogy, the traditional vector space model of Information Retrieval is adapted to image representation model.

This chapter is organized as follows. In Section 4.2, we review the vector space model and how is widely applied to image representation. In Section 4.3, we discussed different term weighting associated with the vector space model. We review different commonly used similarity measures in Section 4.4. Finally, we give a summary and conclusion for this chapter in Section 4.5.

4.2 Vector Space Model

By returning to the text documents, an explicit representation model of the documents is always required to solve most of the information retrieval tasks such as text search, clustering or categorization. Since Shannon showed in 1948 that a wide range of practical problems can be reduced to the problem of estimating the probability distributions of words or n-grams in text [134], the ”bag-of-words” as-

sumption has become a standard practice in text compression, speech recognition, information retrieval and many other applications of Shannon's theory.

Based on the assumption, various data representation models for different information retrieval tasks have been developed, such as Boolean model [7], and fuzzy retrieval model [154]. Nevertheless, most of the recent document clustering, retrieving and classification methods are based on the vector space model [129].

The vector space model represents each text document is represented by a vector where each dimension corresponds to an index term defined as follows.

Definition 3 (Index Term)**.** *The definition of index term in the text documents depends on the application. Typically index terms are single words, keywords, or longer phrases. If the words are chosen to be the index terms, the dimensionality of the vector is the number of words in the vocabulary (the number of distinct words occurring in the corpus).*

If an index term occurs in the document, its value in the vector is non-zero. Several different ways of computing these values, also known as (term) weights, have been developed.

By applying the vector space model to images, each image is described by a set of representative visual keywords, in which each visual keyword is considered to be a visual feature unit (i.e., visterms [69, 122, 120], blobs [26], visual word [173, 72], visual phrase [175, 174, 172, 177]...) called a visual index term. A visual index term is simply assumed to be a visual keyword whose semantic meaning helps in remembering the image's main theme. Thus the index terms can be used to index and summarize the image contents.

Definition 4 (Visual Index Term)**.** *A visual index term is defined as an atomic*

visual representation unit (i.e. visterm [69, 122, 120], blob [26], visual word [173, 72], visual phrase [175, 174, 172, 177]...) in the vector space model, which represents a unique dimension of the vector space.

It is clear that distinct visual index terms of keywords have variable importance in describing image contents. To capture the variable importance of distinct visual index terms, a numerical weight is assigned to each visual index term of an image.

Thus the common framework of vector space model starts with a representation of any image as a feature vector of the weight of the visual index terms that appear in the image collection. In particular, non-binary term-weights (usually tf-idf, term-frequencies and inverse document-frequencies) of the index terms are also contained in the vector [157]. The similarity between two images is computed with a similarity measure on the two corresponding feature vectors.

Let symbols t_i be an visual index term, im_j be an image, and $w_{ij} \geq 0$ be a weight associated with the pair (t_i, im_j). The weight w_{ij} quantifies the importance of the index term ti for describing the image semantic content.

Definition 5 (Vector Space Image Model)**.** *Let k be the number of visual index terms in the image collection, and t_i be a generic visual index term. $T = t_1, t_2, ..., t_k$ is the set of all visual index terms. The weight $w_{ij} \geq 0$ is associated with a visual index term t_i of an image im_j. For a visual index term which does not appear in the image, $w_{ij} = 0$. Then image im_j is represented by a k-dimensional visual index term vector.*

$$\overrightarrow{im_j} = (w_{1,j}, \ ..., \ w_{k,j}) \tag{4.1}$$

The dimensionality of the feature vector is a crucial factor on the efficiency of the corresponding visual representation and hence its scalability. There exist some methods to reduce the dimensionality problem, such as Principle Component Analysis (PCA) [73] [79] and Latent Semantic Index (LSI) model [48]. Krishnapuram et al. [79] proposed an approach for reducing a 500-dimensional problem to a 10-dimensional problem by using the PCA method. The LSI model [48] introduces an interesting conceptualization of the information retrieval problem based on the theory of singular value decomposition. The main idea of LSI model is to map each document or image and the query vector into a lower dimensional space which is associated with the concepts. Thus the efficiency of the algorithm in the reduced space might be superior to the same algorithm in the space of original index terms.

Baeza-Yates et al. [7] summarizes the main advantages of vector space model as follows:

1. Its term weighting scheme improves retrieval performance.
2. Its partial matching strategy allows the retrieval of documents that approximate the query conditions.
3. Its cosine similarity measure ranks the documents according to their relevance degree of similarity to the query.

4.3 Term weighting measures

Since term weighting is a key technique in IR [128, 6], recent research on image representation explores its use in weighting visual index terms. There is much research on term weighting techniques with little consensus on which method is

the best. Two major factors in term weighting are tf (term frequency) and idf (inverse document frequency). A third factor is the normalization factor, which converts a feature into unit-length vector to eliminate the difference between short and long documents. Many visual representation methods use weighting schemes based on these factors [120, 173, 72, 175, 174].

This section gives a brief review for several commonly used term weighting measures associated with the vector space model.

4.3.1 Term Frequency (tf)

Index terms that are frequently mentioned in individual documents, or document excerpts, appear to be useful as recall enhancing devices. This suggests that a *term frequency (tf)* factor be used as part of the term-weighting system measuring the frequency of occurrence of the terms in the document or query texts. For a document, the set of weights determined by *tf* may be viewed as a quantitative digest of that document.

In this view of a document, the exact ordering of the terms in a document is ignored but the number of occurrences of each term is material. We only retain information on the number of occurrences of each term. Thus, the document 'Mary is quicker than John' is, in this view, identical to the document 'John is quicker than Mary'. Nevertheless, it seems intuitive that two documents with similar representations based on *tf* are similar in content. Term-frequency weights have been used for many years in automatic indexing environments [94, 157, 7].

4.3.2 Inverse Document Frequency (idf)

The inverse document frequency (idf) was proposed by Spärck Jones [74] *four* decades ago and has become one of the popular measures for representing the importance of index terms in a text document corpus. The idf is defined as the logarithm of the ratio of the number of documents (N) in a collection to the number of documents containing the given term (df). This means that idf is measuring the ability of a term to discriminate the subject or topic of the documents. In a set of documents, rare terms have higher idf values and common terms have lower values. The idf is often described as a heuristic without a theoretical basis, thus it becomes a magnet attracting many researchers to explore the theoretical principle for explaining why it can work so well. Stephen's paper [125] gives a review for some of these attempts. However, these contributions are achieved under some limited assumptions. So far, the idf is still a theoretical heuristic.

The work of Papineni [116] makes some appeals to information theory, such as maximum entropy, Kullback-Leibler distance, and mutual information. Based on the agreement that the idf is a heuristic, their work shows that the idf is optimal in the precision sense of information retrieval. They first consider information retrieval as a classification problem with each document in the collection being a class. They then build a classifier that scores the documents given a query. To train the classifier, they treat each document as a query that retrieves itself. Thus the classifier they develop is an exponential model similar to the one in the maximum entropy framework, but without the usual normalization. In the case when there is a single binary feature in this model, the optimal solution is

stunningly simple in contrast to the solution in the regular conditional maximum entropy framework. The single feature of a word that examines the occurrence of the word in both the query and the document is exactly a binary feature. Thus idf is the optimal weight of this feature of word in document self-retrieval.

The idf is applied to images in the same manner as tf and is defined as the logarithm of the ratio of number of images (N) in a dataset to the number of images containing the given visual index term (df).

4.3.3 The $tf \times idf$ weighting scheme

The $tf \times idf$ [125] is the most famous weighting scheme which is widely used in modern information retrieval. The $tf \times idf$ weight [125]comprises two components: a term frequency tf that provides a local measure for the feature term associated with the document, and an inverse document frequency idf that provides a global measure for the term among the documents in the document collection or corpus. Given $tf \times idf$ weighting scheme assigns to term t a weight in document d given by:

$$tf - idf_{t,d} = tf_{td} \times idf_t \tag{4.2}$$

In other words, $tf - idf_{t,d}$, assigns to an index term t a weight in document d that is

- higher when t occurs many times within a small number of documents (thus lending high discriminating power to those documents);
- lower when the term occurs fewer times in a document, or occurs in many documents (thus offering a less pronounced relevance signal);
- lower when the term occurs in most of all documents.

The $tf \times idf$ weighting scheme has been proven to be extremely robust and difficult to beat, even by much more carefully worked out models and theories [125].

4.3.4 Variant normalized term weighting measures

For assigning a weight for each index term in each document, a number of alternatives to tf and $tf \times idf$ have been considered. We discuss some of the principal ones here.

4.3.4.1 Sublinear tf scaling

It seems unlikely that twenty occurrences of an index term in a document truly carry twenty times the significance of a single occurrence. Accordingly, there has been considerable research into variants of term frequency that go beyond counting the number of occurrences of a term. A common modification is to use instead the logarithm of the term frequency, which assigns a weight given by:

$$wf_{t,d} = \begin{cases} 1 + \log tf_{t,d} & if\ tf_{t,d} > 0 \\ 0 & otherwise \end{cases} \tag{4.3}$$

In this form, we may replace tf with some other function wf as in (4.2), to obtain:

$$wf - idf_{t,d} = wf_{t,d} \times idf_{t,d} \tag{4.4}$$

4.3.4.2 Maximum tf normalization

One well-studied technique is to normalize the tf weights of all terms occurring in a document by the maximum tf in that document. For each document d, let

$tf_{max(d)} = max_{(\tau \in d)} tf_{\tau,d}$ where τ ranges over all terms in d. Then, we compute a normalized term frequency for each term t in document d by:

$$ntf_{t,d} = \alpha + (1 + \alpha)\frac{tf_{t,d}}{tf_{max(d)}} \tag{4.5}$$

where α is a value between 0 and 1 and is generally set to 0.4, although some smoothing early work used the value 0.5. The term α in (4.5) is a smoothing term whose role is to damp the contribution of the second term-which may be viewed as a scaling down of tf by the largest tf value in d. The main idea of $maximumtfnormalization$ is to mitigate the following anomaly: we observe higher term frequencies in longer documents, merely because longer documents tend to repeat the same words over and over again. To appreciate this, consider the following extreme example: suppose we were to take a document d and create a new document d' by simply appending a copy of d to itself. While d' should be no more relevant to any query than d is, the use of (4.2) would assign it twice as high a score as d. Replacing $tf \times idf_{(t,d)}$ in (4.2) by $ntf \times idf_{(t,d)}$ eliminates the anomaly in this example. Maximum tf normalization does suffer from the following issues:

- The method is unstable in the following sense: a change in the stop word list can dramatically alter term weightings (and therefore ranking). Thus, it is hard to tune.
- A document may contain an outlier term with an unusually large number of occurrences of that term, not representative of the content of that document.
- More generally, a document in which the most frequent term appears roughly as often as many other terms should be treated differently from

one with a more skewed distribution.

4.4 Similarity measures

The essential property of a similarity measure is to accurately reject the degree of similarity between two data objects. It's widely recognized that similarity measure is a key factor in many data analysis applications due to different feature types, different cluster shapes, and different clustering principles. There is a large number of similarity metrics reported in the related literature; only the most common measures are reviewed in this section.

In general, similarity measures are ranged in the interval of $[0, 1]$, with 1 denoting the highest similarity and 0 being the lowest similarity. Similarity measures satisfy reflexive, symmetric properties.

- Reflexivity: $\forall x,\ sim(x, x) = 1$
- Symmetry: $\forall x, y\ sim(x, y) = sim(y, x)$

Reflexivity means that each object has maximum similarity to itself. Symmetry ensures that the similarity between two objects is independent to the direction of comparison.

In many data analysis applications, similarity measures are replaced with distances, sometimes dissimilarities as well. The distance between two instances is a non-negative number, with 0 representing the shortest distance. Distance satisfies reflexive and symmetry properties. In addition, distance also satisfies the triangle inequality.

- Triangle inequality: $\forall x, y, z\ d(x, y) \leq d(x, z) + d(y, z)$

There are many ways to transfer a distance measure to a similarity measure.

For example, $sim(x, y) = \frac{d_{max} - d_{(x,y)}}{d_{max} - d_{min}}$, where d_{max} is the maximum value of distances and d_{min} is the minimum value of distances.

4.4.1 Minkowski distance

As we mentioned before, in the classical vector space image model, an image is represented as a vector. Each dimension corresponds to a separate weight for the visual index term. If a visual index term occurs in a image, its value in the vector is non-zero.

A commonly used class of distance functions is known as the family of Minkowski distance.

Let two feature vectors x, y represent two objects, the Minkowski distance of them is given by:

$$d(x, y) = \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p} \tag{4.6}$$

Three special cases of Minkowski distance are:

- $p = 1$: Hamming Distance

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i| \tag{4.7}$$

- $p = 2$: Euclidean Distance

$$d(x, y) = \sqrt[2]{\sum_{i=1}^{n} |x_i - y_i|^2} \tag{4.8}$$

– $p = \infty$: Tschebyshev Distance

$$d(x, y) = max_{i=1,2,...,n} |x_i - y_i| \tag{4.9}$$

4.4.2 Cosine similarity measure

The most commonly used similarity measure in information retrieval, indexing, relevance ranking is the cosine correlation measure [127], which is defined as:

$$d(x, y) = \frac{\sum_n^{i=1} x_i y_i}{\sqrt{\sum_n^{i=1} x_i^2 \sum_n^{i=1} y_i^2}} \tag{4.10}$$

The cosine similarity measure computes the similarity or relevancy of two images by comparing the deviation of angles between the two image vectors.

4.4.3 Jaccard similarity measure

Another commonly used similarity measure is the Jaccard similarity measure (also called Jaccard similarity coefficient) [66], which is defined as follows:

$$d(x, y) = \frac{\sum_n^{i=1} x_i y_i}{\sum_n^{i=1} x_i^2 + \sum_n^{i=1} y_i^2 - \sum_n^{i=1} x_i y_i} \tag{4.11}$$

In the case of binary feature vector it is simplified to be:

$$d(x, y) = \frac{|x \cap y|}{|x \cup y|} \tag{4.12}$$

Hence, Jaccard similarity measure is defined as the size of the intersection divided by the size of the union of the sample sets.

4.5 Summary and conclusion

This chapter introduces a brief review of a vector space model that represents a document with a vector that captures the relative importance of the terms in a document. The vector space model is fundamental to a number of operations ranging from scoring documents on a query, document classification and document clustering. This vector space model is adapted to image representation, and is used for similarity matching, retrieval, classification and ranking of images. In our approach, the Vector Space Model is applied to different levels of the proposed visual representation, and used for similarity matching and retrieval of images.

In addition, a number of term-weighting measures are described in these chapters which are combinations of term frequency, collection frequency, and normalization components. The $tf \times idf$ weighting scheme has been proven to be extremely robust by comparing to others [125]. There exist also a large number of similarity measures reported in the related literature; only the most common measures that are associated with vector space model are reviewed in this chapter.

The best weighting scheme in IR does not guarantee good performance in image representation since most of the weighting schemes do not consider the spatial location of the local patches. However, such information in an image is important for classifying and retrieving the images [170, 70]. For this reason, we create a weighting scheme that weights for the first level of the proposed representation (SSIVW) according to the spatial constitution of an image content rather than

the number of occurrences. Besides, we have used the $tf \times idf$ weighting for the higher-level of representation (SSIVP) where the spatial constitution is not considered.

Part II

A Semantic Higher-Level Visual Representation

Chapter 5

Enhanced Bag of Visual Words

Contents

5.1 Overview

As mentioned in the Introduction, we make use of several chains of processes from the lower level to a higher level of representation, in order to extract and refine the information incrementally. With an analogy between text and image documents, we consider that an image is composed of visual words and it can be represented as a *Bag of Visual Words*.

The objective of this chapter is to introduce the different hierarchical process (see Figure 5.1) that are performed in order to enhance the classical approach of bag of visual words (BOW).

This chapter is organized as follows. In Section 5.2, we present the Fast-Hessian detector [10]. In Section 5.3, we introduce the canny edge detector with the Sobel operator [24]. A Vector Median Filter (VMF) [150] is applied to remove the color noise as described in Section 5.4. We model the color and position feature space for set of interest and edge point based on the Gaussian Mixture Model (GMM) [15] in Section 5.5. In Section 5.6, we describe how to extract and describe local features. In addition to SURF [10], we introduce a new local feature extractor, the Edge context, which plays the role of complimentary descriptor to SURF descriptor. It describes, at each interest point, the distribution of the edge points that are in the same Gaussian cluster by returning to the 5D color-spatial space. In Section 5.7, the quantization of the merged features into visual words is described based on Hierarchical Agglomerative Clustering (HAC) and repeated k-means clustering that hierarchically partition the feature space in order to build

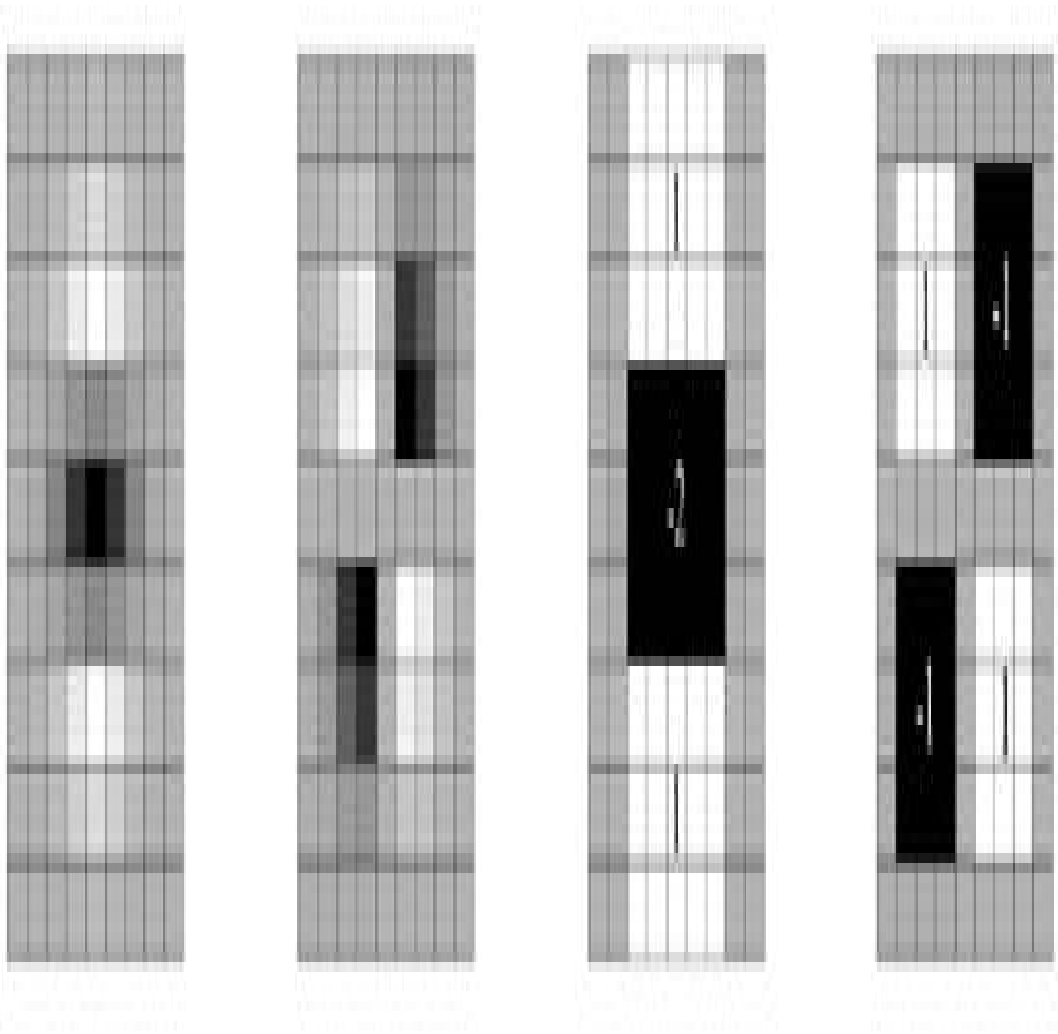

Figure 5.1: Overview of the different processes for generating the visual words.

a visual word vocabulary tree. Finally, we give a summary and conclusion for this chapter in Section 5.8.

5.2 Interest points detection

We use the *Fast-Hessian* detector [10] to detect interest points. This detector is based on the Hessian matrix because of its good performance in computation time and accuracy. However, rather than using a different measure for selecting the location and the scale (as it is done with the Hessian-Laplace detector [104]), it relies on the determinant of the Hessian for both. Given a point $X = (x, y)$ in an image I, the Hessian matrix $H(X, \sigma)$ at scale is defined as follows:

$$\begin{bmatrix} L_{xx}(X,\sigma) & L_{xy}(X,\sigma) \\ L_{xy}(X,\sigma) & L_{yy}(X,\sigma) \end{bmatrix} \tag{5.1}$$

where $L_{xx}(X, \sigma)$ is the convolution of the Gaussian second order derivative $H(x, \sigma)$ with the image I in point X, and similarly for $L_{xy}(X, \sigma)$ and $L_{xy}(X, \sigma)$.

Gaussians are optimal for scale-space analysis, as shown in [77]. In practice, however, the Gaussian needs to be discretized and cropped (Figure 5.2, (a) and (b), and even with Gaussian filters same aliasing still occurs when the resulting images are sub-sampled. Given Lowe's success with LoG approximations [93], Bay et al. use box filters (Fig. 5.2, (c) and (d)) for approximation. These box filters approximate the second order Gaussian derivatives, and can be evaluated very fast using *integral images*, independently from the size.

The 9×9 box filters in Figure 5.2 are approximations for Gaussian second order derivatives with $\sigma = 1.2$, and they represent the lowest scale (i.e. highest spatial resolution) as proposed by Bay et al. These approximations are denoted by D_{xx}, D_{yy}, D_{xy}. The weights applied to the rectangular regions are kept simple for computational efficiency with additional further balance to the relative weights

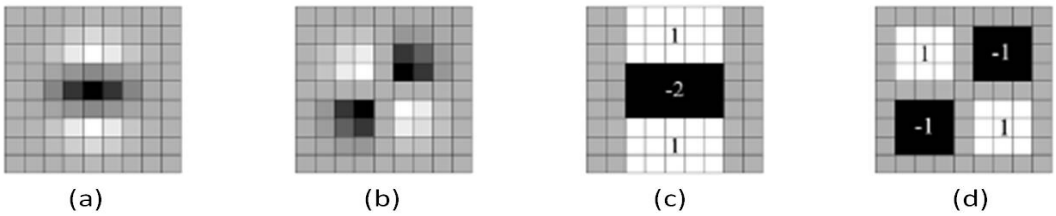

Figure 5.2: The (discretized and cropped) Gaussian second order partial derivatives in y-direction (a) and xy-direction (b), and the approximations thereof using box filters (c) and (d). The gray regions are equal to zero [10].

in the expression for the Hessian's determinant.

The scale space is analyzed by up scaling the filter size rather than iteratively reducing the image size. The output of the above 9×9 box filters is considered as the initial scale layer, to which is referred as the scale $s = 1.2$ (corresponding to Gaussian derivatives with $\sigma = 1.2$). The following layers are obtained by filtering the image with gradually bigger masks, taking into account the discrete nature of integral images and the specific structure of these filters. This results in filters of size 9×9, 15×15, 21×21, 27×27, etc.

Finally, in order to localize interest points in the image and over scales, a non-maximum suppression in a $3 \times 3 \times 3$ neighborhood is applied. The maxima of the determinant of the Hessian matrix are then interpolated in scale and image space with the method proposed by Brown et al. [18]. Scale space interpolation is especially important in this case, as the difference in scale between the first layers of every octave is relatively large. Figure 5.3 shows some examples of detected interest points using the *Fast-Hessian* detector.

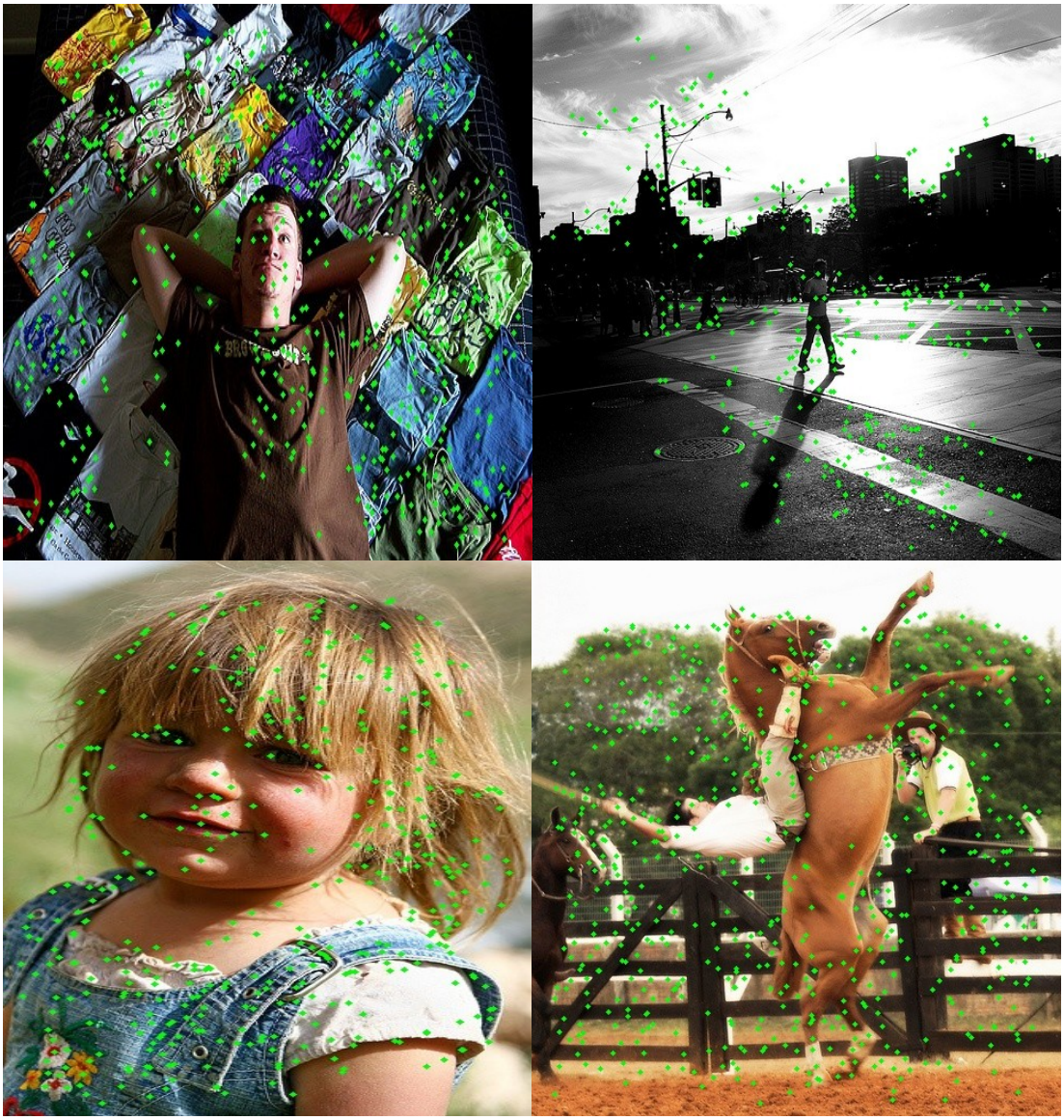

Figure 5.3: Examples of detected interest points using the *Fast-Hessian* detector

5.3 Edge points detection

The *Canny* algorithm [24] is used to detect edge points. We used the Canny algorithm since it is adaptable to various environments. Its parameters allow it to be tailored to the recognition of edges of different characteristics from different

image databases.

This algorithm makes use of several operator widths to cope with varying image signal-to-noise ratios, and operator outputs were combined using a method called **feature synthesis**, where the responses of the smaller operators are used to predict the large operator responses. If the actual large operator outputs differ significantly from the predicted values, some new edge points are marked. It is therefore possible to describe edges that occur at different scales, even if they are spatially coincident.

In our approach, we use the *Sobel* operator as an edge detection operator within the *Canny* algorithm to detect horizontal, vertical and diagonal edges in the image. The *Sobel* operator returns a value for the first derivative in the horizontal direction (G_y) and the vertical direction (G_x). From this, the edge gradient and direction can be determined:

$$G = \sqrt{G_x^2 + G_y^2} \tag{5.2}$$

$$\Theta = \frac{G_x}{G_y} \tag{5.3}$$

The edge direction angle is rounded to one of four angles representing vertical, horizontal and the two diagonals (0, 45, 90 and 135 degrees for example). Figure 5.4 shows examples of the detected edge points.

Figure 5.4: Examples of edge points detected by *Canny* algorithm with *Sobel* operator

5.4 Color filtering using Vector Median Filter (VMF)

Color images are often corrupted with noise due to capturing devices or sensors. It significantly improves visual representation accuracy if an effective filter is applied to remove the color noise. The pre-process can be essential, especially when the image retrieval or classification results are used for human interpretation.

Since each individual channel of a color image can be considered a monochrome image, traditional image filtering techniques often involve the application of scalar filters on each channel separately. However, this disrupts the correlation that exists between the color components of natural images represented in a correlated color space, such as sRGB [150]. Since each processing step is usually accompanied by a certain inaccuracy, the formation of the output color vector from the separately processed color components usually produces color artifacts.

Thus, vector filtering techniques that treat the color image as a vector field are more appropriate. With this approach, the filter output $\hat{x}(N+1)/2$ is a function of the vectorial inputs $x_1, x_2, ..., x_N$ located within the supporting window W(see Figure 5.5).

Assuming an RGB color image x, each pixel $x_i = [x_{i1}, x_{i2}, x_{i3}]^T$ represents a three-component vector in a color space as shown in Figure 5.6.

The color image x is a vector array or a two-dimensional (2D) matrix of the three-component samples x_i with x_{ik} denoting the $R(k = 1), G(k = 2)$, or B component $(k = 3)$ (see Figure 5.7) .

We used the Vector Median Filter (VMF), which is the most popular vector filter [5]. The VMF is a vector processing operator that has been introduced

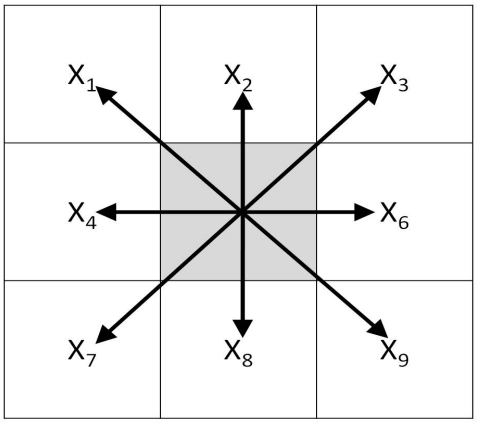

Figure 5.5: The 3×3 filtering mask with the window center $x(N+1)/2 = x_5$.

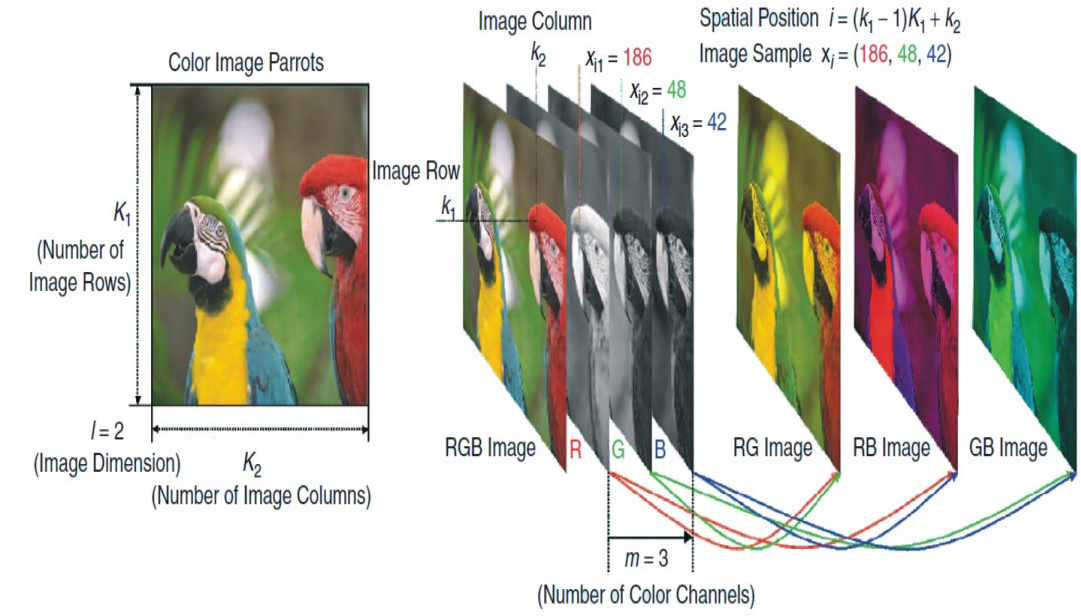

Figure 5.6: Color image representation in the RGB color domain [150]

as an extension of the scalar median filter. The generalized Minkowski metric $\|x_i - x_j\|_L$ is used to quantify the distance between two color pixels x_i and x_j in the magnitude domain. To speed up the calculation of the distances between the

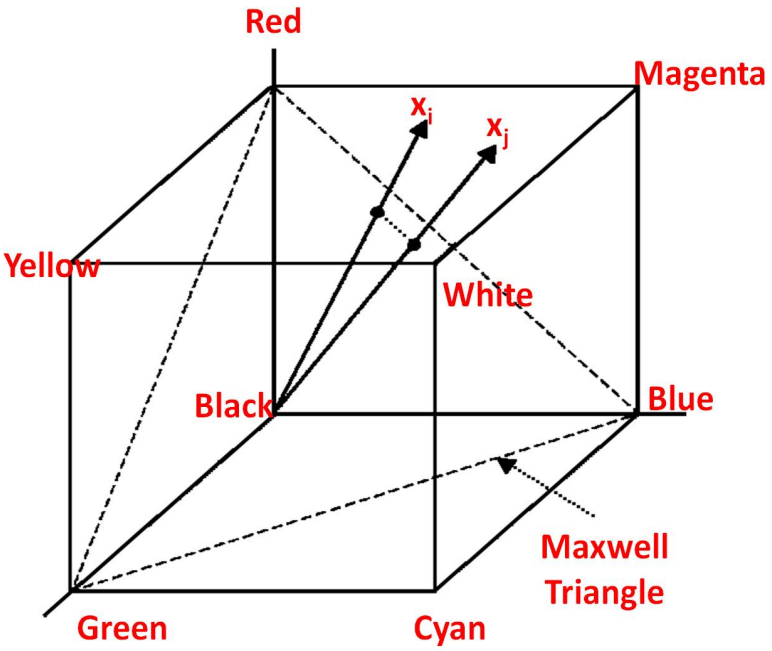

Figure 5.7: RGB color Cube

color vectors, we used VMF based on the linear approximation of the Euclidean norm as proposed by [9].

The VMF output is the sample $x_{(1)} \in W$ that minimizes the distance to the other samples inside W. Since the ordering can be used to determine the positions of the different input vectors without any a priori information regarding the signal distributions, vector order-statistics filters, such as the VMF, are robust estimators [9].

5.5 Gaussian Mixture Model (GMM) for the color-spatial feature space

Based on the Gaussian Mixture Model (GMM) [15], we model the color and position feature space for set of interest and edge points. This model is used to extract the Edge Context descriptor and later for spatial weighting [41, 42].

A $5D$ color-spatial feature vector is with the 3 dimensions for RGB color, plus 2 dimensions (x, y) for the position, in order to represent each interest and edge point. In an image with m interest/edge points, a total of m $5D$ color-spatial feature vectors: $f_1 ... f_m$ can be extracted.

The set of points is assumed to be a mixture of n Gaussians in the $5D$ color-spatial feature space and the Expectation-Maximization (EM) [35] algorithm is used to iteratively estimate the parameter set of the Gaussians.

The parameter set of the Gaussian mixture is $\theta = \{\mu_i,\ \Sigma_i,\ p_i\}_{i=1}^{n}$ where :

- μ_i is the mean of the i$^{\text{th}}$ Gaussian cluster.
- Σ_i denotes the covariance matrix.
- p_i represents the prior probability of the i$^{\text{th}}$ Gaussian cluster.

By applying Bayes theorem at each E-step, we estimate the probability of a particular feature vector f_j belonging to the i^{th} Gaussian according to the outcomes from the last M-step as follows.

$$P(g_i|f_j, \theta_t) = \frac{P(f_j|g_i, \theta_t)P(g_i|\theta_t)}{P(f_j)} \tag{5.4}$$

$$P(f_j) = \sum_{k=1}^{n} P(f_j|g_k, \theta_t)P(g_k|\theta_t) \tag{5.5}$$

Where g_i denotes the Gaussian which f_j comes from and θ_t is the parameter set at the t^{th} iteration.

At each M-step, the parameter set of the n Gaussians is updated toward maximizing the log-likelihood, which is:

$$Q(\theta) = \sum_{j=1}^{m} \sum_{i=1}^{n} P(g_i|f_j, \theta_t) ln(P(f_j|g_i, \theta_t) P(g_i|\theta_t)) \tag{5.6}$$

When the algorithm converges, the parameter sets of n Gaussians as well as the probability $P(g_i|f_j)$ are obtained. For each feature vector f_j, we indicate the most likely Gaussian cluster to which it belongs as follows.

$$P_{f_j}^{max} = argmax_{g_i}(P(g_i|f_j)) \tag{5.7}$$

Finally, the set of interest and edge points in an image can be grouped into n Gaussian clusters according to the Gaussian where their $5D$ feature vectors belong to.

5.6 Extracting and describing local features

In our approach, we use the SURF low-level feature descriptor that describes how the pixel intensities are distributed within a scale-dependent neighborhood of each interest point detected by the Fast-Hessian.

In addition to the SURF descriptor, we introduce a novel *Edge Context*descriptor at each interest point detected by the Fast-Hessian, based on the distribution of the edge points in the same Gaussian (by returning to the 5D color-spatial feature space described in Section 5.5).

5.6.1 SURF

This descriptor is similar to SIFT [93], but Bay et al. have used integral images [159] in conjunction with filters known as Haar wavelets in order to increase the robustness and to decrease the computation time. Haar wavelets are simple filters that can be used to find gradients in the x and y directions. The extraction of the descriptor can be divided into different distinct steps.

1. A reproducible *orientation* for the interest points are identified in order to be invariant to rotation. For this purpose, the Haar-wavelet responses in x and y direction, as shown in Figure 5.8, and this in a circular neighborhood of radius $6s$ around the interest point, where s is the scale at which the interest point is detected. Once the wavelet responses are calculated and weighted with a Gaussian ($\sigma = 2.5s$) centered at the interest point, the responses are represented as vectors in a space with the horizontal response strength along the abscissa and the vertical response strength along the ordinate. The dominant orientation is estimated by calculating the sum of all responses within a sliding orientation window covering an angle of $\pi/3$. The horizontal and the vertical responses within the window are summed.

Figure 5.8: Haar wavelet types used for SURF.

2. The square regions centered on each interest point are generated, and turned along the orientation at the interest point. The size of this window is

20 times as big as the scale of the detected interest point and the region is split up regularly into smaller 4×4 square sub-regions. Figure 5.9 shows examples of images after SURF extraction. The green points resemble the interest points where the blue square represent the descriptor windows around each point. In addition, the inclined green lines from each interest point to the border of contained descriptor window represent the orientation at these points.

3. Haar wavelet response in horizontal direction (dx) and in vertical direction (dy) are summed up over each sub-region and form a first set of entries to the feature vector. In order to bring in the information about the polarity of the intensity changes, the sum of the absolute values of the dx and dy responses are also extracted. Hence, each sub-region has a four-dimensional descriptor vector v for its underlying intensity structure as follows.

$$v = (\sum dx, \sum dy, \sum |dx|, \sum |dy|) \tag{5.8}$$

This results in a descriptor vector for all 4×4 sub-regions of length 64. The wavelet responses are invariant to a bias in illumination (offset). Invariance to contrast is achieved by turning the descriptor into a unit vector.

5.6.2 A new low level feature (Edge Context)

Intensity-based descriptors make more direct use of pixel intensity values [89]. This turns out to be quite a challenge since these techniques do not cope well with the large distortions that must be handled due to pose and illumination variations. Moreover, all other information like shape and color are ignored in

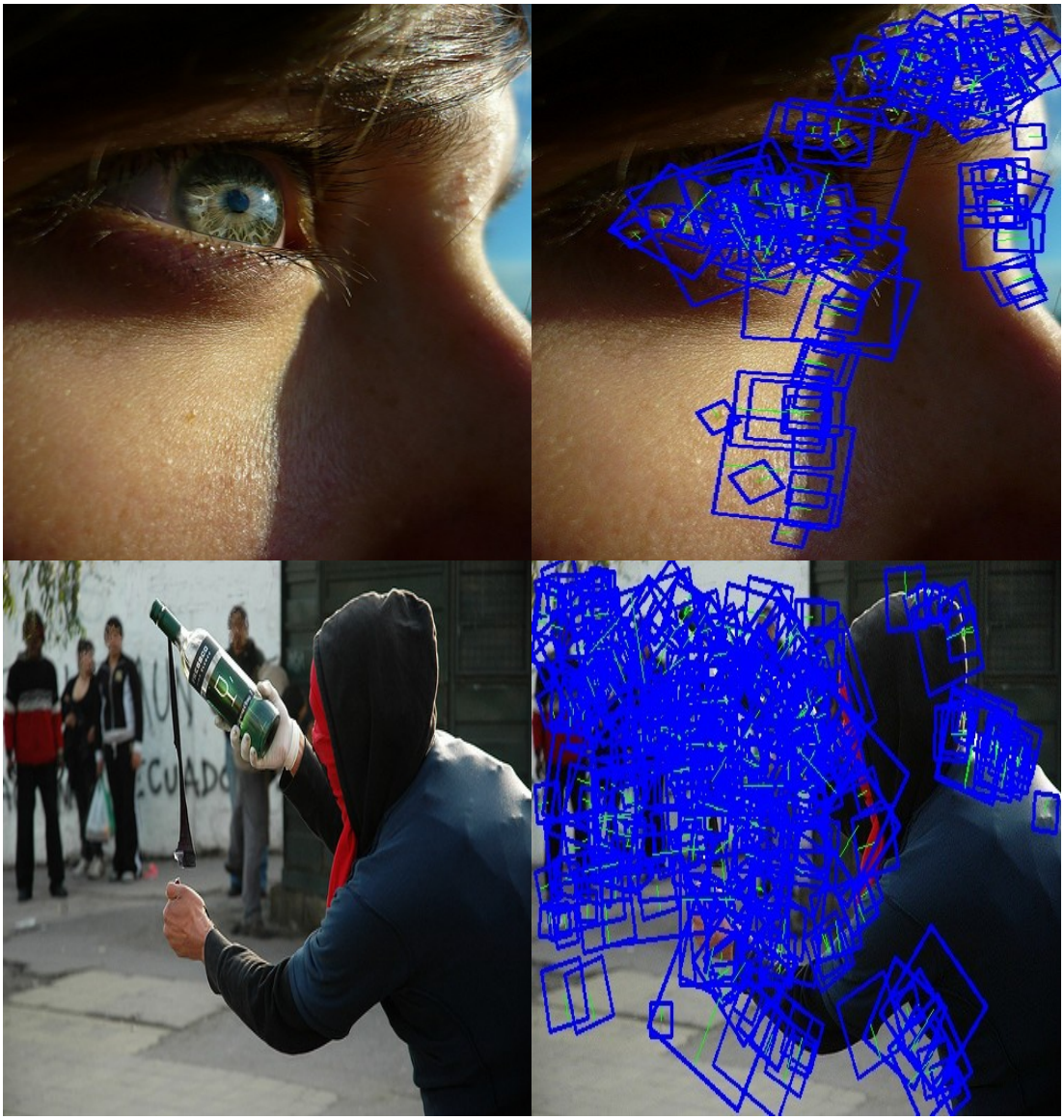

Figure 5.9: Examples of SURF descriptor windows at different interest points.

these descriptors, however such information is essential for many tasks such as handwritten digit recognition [82, 23], face recognition [107], and isolated 3D object recognition [112].

As one of the contributions in our work, we propose a novel descriptor, the

Edge Context [40, 41], that could play a role as a complementary descriptor in addition to the SURF descriptor.

5.6.2.1 Descriptor components

As shown in Figure 5.10, vectors from each interest point in the 2D spatial image space are drawn to all other edge points that are within the same Gaussian cluster in 5D color-spatial feature space. The Edge Context descriptor for each interest point is represented as a $2D$ histogram as follows.

- m horizontal bins for the different values of r (*magnitude* of the drawn vector from the interest point to the edge points).
- n vertical bins for θ (*orientation angle* of the drawn vector from the interest point to the edge points).

The histogram data contains the frequency of the edge points that fall in each *bin* of the grid defined by r and θ values.

The number of bins of the histogram can be chosen. In our work, the 6 bins for r and 4 bins for θ provided the best results according to the experimental results in Chapter 8.

This descriptor is inspired by the shape context descriptor proposed by Belongie et al. [13] with respect to the extracted information from edge point distribution. Describing the distribution of these points enriches the descriptor with more information, rather than the intensity described by SURF. Moreover, the distribution over relative positions is a robust, compact, and highly discriminative descriptor.

Figure 5.10: Examples of vectors (white lines) drawn from a given interest point (green dot inside the red circle) to all other edge points (black points) that are in the same 5D Gaussian cluster as the interest point.

5.6.2.2 Invariance and robustness

A low-level descriptor should be invariant to scaling and translation changes, and robust under small affine transformations, occlusion and presence of outliers.

In certain applications, one even needs a complete invariance to rotation. For this novel descriptor, many of these invariance requirements are satisfied as follows.

- The invariance to translation is intrinsic to the Edge Context since the distribution of the edge points is measured with respect to fixed interest point.
- Invariance to scale is achieved by normalizing the radial distance by a mean distance between the whole set of points inside a given Gaussian in the 5D color-spatial feature space.
- The descriptor can provide complete rotation invariance if this is desirable. Instead of using the absolute coordinate frame for computing the Edge Context at each interest point, one can use the tangent vector at each point as the positive x-axis. In this way, the reference frame turns with the tangent angle, therefore the result is a completely rotation invariant descriptor.

5.6.3 Fusion of the Edge Context and the SURF descriptors

Following the visual construction part in Figure 5.1, after extracting the Edge Context descriptor, a fusion with SURF descriptor is performed. This merged feature vector is composed of 88 dimensions (64 from SURF + 24 from the Edge Context descriptor). Hence, the new feature vector contains the information on the distribution of the intensity and the distribution of the edge points in the $5D$ spatial-color space. It enriches the image representation with more local information.

5.7 Local features quantization

The quantization of the merged feature vectors (SURF + Edge Context feature vector) is performed in order to construct a visual word vocabulary tree similar to [113]. The visual word vocabulary tree is computed using a Divisive Hierarchical K-Means clustering that hierarchically partitions the feature space. In addition to Divisive Hierarchical K-Means, we used group-average Hierarchical Agglomerative Clustering (HAC) to bootstrap K-means in order to avoid problems of bad seed selection [41].

5.7.1 Initial seeds of the quantization cells using Hierarchical Agglomerative Clustering (HAC)

Let n be the number of feature vectors to be quantized, and k denotes the number of children of each node of the tree [1].

A random subset of size $\sqrt{n}$ is sampled from the entire set of the vectors, and the group-average HAC is run on this subset.

HAC considers each point in the feature space as a separate cluster, and combines the clusters with the maximum similarity. The similarity between clusters is measured as a group average. When the required number of clusters k is reached, the algorithm is stopped. Figure 5.11 shows an example of a clustering obtained by cutting the *dendrogram* at a desired level. The overall algorithm complexity is $\theta(n)$, and it avoids the problems of bad seed selection since no more any random initialization is needed, as for K-Means.

1. Note that here, k does not refer to the traditional final number of clusters.

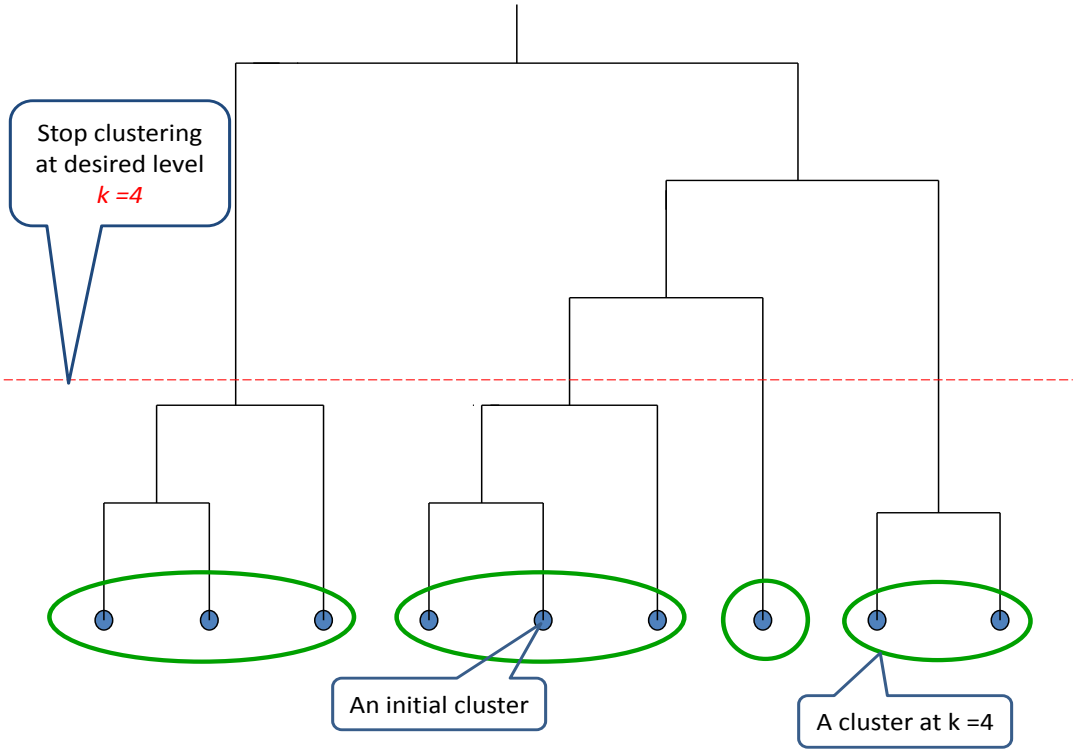

Figure 5.11: An example of a HAC dendrorgram cut at a desired level.

5.7.2 Visual word vocabulary tree construction using Divisive Hierarchical K-Means Clustering

$K-Means$ process is run on the initial seeds that are obtained from the HAC, recursively defining quantization cells by splitting each quantization cell into k new parts. The tree is determined level by level, up to some maximum number of levels L, and each division into k parts is only defined by the distribution of the fused feature vectors that belong to the parent quantization cell.

In the online phase, each merged feature vector is simply propagated down the tree by comparing at each level the feature vector to the k candidate cluster centroids (represented by k children in the tree) and choosing the closest one.

Finally, each vector is mapped to its closest visual word index. Figure 5.12 shows an example of a merged feature vector assigned into a discrete visual word (index 6).

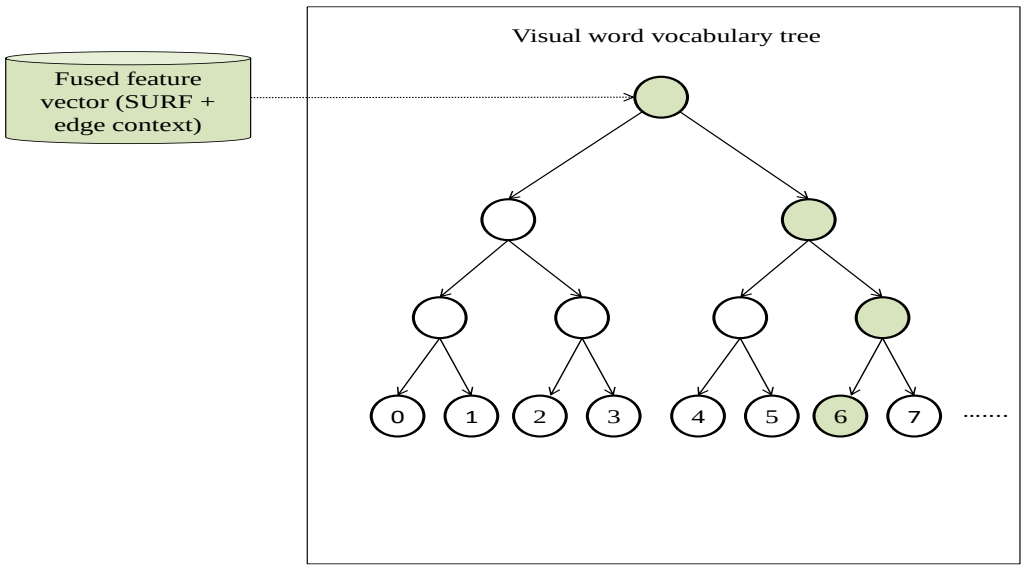

Figure 5.12: Example of assigning a merged feature vector into a discrete visual word.

This is a simple matter of performing k dot products at each level, resulting in a final vocabulary size is $K = k \times L$.

Note that the tree directly defines both the visual word vocabulary and an efficient search procedure in an integrated manner. This is different from for example defining a visual word vocabulary non-hierarchically, and then devising an approximate nearest neighbor search in order to find the visual words efficiently. This hierarchical approach overcomes two major problems related to traditional direct K-means clustering because of the following:

- The clustering is more efficient during learning step.
- The mapping of visual features to discrete visual word is much faster than

using a plain list of visual words.

5.8 Summary and conclusion

This chapter introduces an enhanced approach to construct the bag of visual words (BOW) representation, which is the first level of representation in the proposed approach. We introduced a hierarchal approach in order to enhance the classical bag of visual words in different aspects.

First, we detect the interest and the edge points using the Fast Hessian and the canny edge detector with the Sobel operator respectively. Second, the image noise is filtered using the Vector Median Filter (VMF), which is a pre-step before modeling the 5D color-spatial feature space for the set of interest and edge points based on the Gaussian Mixture Model (GMM). Third, we extract SURF local features at each interest point. In addition to SURF feature, we establish a new local feature descriptor, the Edge Context, which plays a role of a descriptor complimentary to SURF descriptor. It describes at each interest point the distribution of the edge points that are in the same Gaussian cluster by returning to the 5D color-spatial space. Finally, the two local feature vectors (SURF +Edge Context) are merged to get final local feature vectors. The quantization of the merged features into visual words (VWs) is achieved by two clustering steps. A hierarchical agglomerative clustering is performed to overcome the problem of the initial seed for the repeated K-Means clustering that hierarchically partition the local feature space. Finally, a visual word vocabulary tree is built.

Chapter 6

Multilayer Semantic Significance Analysis (MSSA) Model

Contents

6.1 Overview

In the previous chapter, we have introduced the different processes to generate the visual words. One of these processes is the feature quantization that generates many unnecessary and insignificant visual words, which are noisy in retrieval and classification.

In this chapter, a new multilayer semantic significance analysis (MSSA) model is introduced in order to study the semantic inferences of the constructed visual words, based on their probability distributions regarding to the relevant visual latent topics [39]. The estimation of the semantic inference of the visual words is important in our approach in order to select semantically significant visual words and eliminate the insignificant visual words. In addition, this model is useful to study the semantic inference of different atomic visual representation unit (such as visterm [69, 122, 120] and visual phrases [175, 174, 172, 177])

This chapter is organized as follows. In Section 6.2, we discuss the motivation of proposing this multilayer probabilistic model, MSSA, based on two different latent topic layers (*high latent topics* and *visual latent topics*). The different generative processes that model the probabilistic distribution of different elements in the MSSA model are presented in Section 6.3. In Section 6.4, the KKT conditions are used [80] to derive new multiplicative update rules in order to estimate the parameters of the MSSA model. In Section 6.5, the number of latent topics in the MSSA model is estimated using three different mode selection criteria: the Akaike information criterion (AIC), the Bayesian information criterion (BIC), and the Minimum Description Length (MDL).

6.2 Motivation

In image representation, capturing the essential statistical characteristics of different visual representation units allows to build a relevant representation for images, which is often more parsimonious and less noisy. Especially in the BOW image representation, the vocabulary creation process, based on clustering algorithms such as K-Means, is quite rude and can lead to many noisy visual words. Such visual words add ambiguity in the image representation. Thus, it reduces the effectiveness of the visual representation in retrieval or classification, and a statistical criterion is needed to study the semantic significance of the constructed visual words.

Among the existing methods that extract statistical characteristics, the probabilistic topic models play an important role. Probabilistic topic models extract a set of latent topics from a corpus, and therefore they represent the images in a new latent semantic space.

Many of these models introduce only one latent topic layer between the documents (images) and the representation atomic unit (i.e., visual words). However, in our understanding, every image is assumed to consist of one or more visual aspects, which in turn are combined into the higher-level aspects. This is very natural since images consist of multiple objects or scenes, which belong to different categories or classes. Figure 6.1 shows an example of different high and visual aspects in some images. In this figure, the *face* can be a visual aspect and the *person* can be the high aspect.

A new probabilistic topic model is designed to take in consideration the hierarchal consistence of the image, without adding much complexity in the process

of the parameters initialization and estimation. We introduce the new multi-layer probabilistic topic model (MSSA) that considers the different aspects of the image.

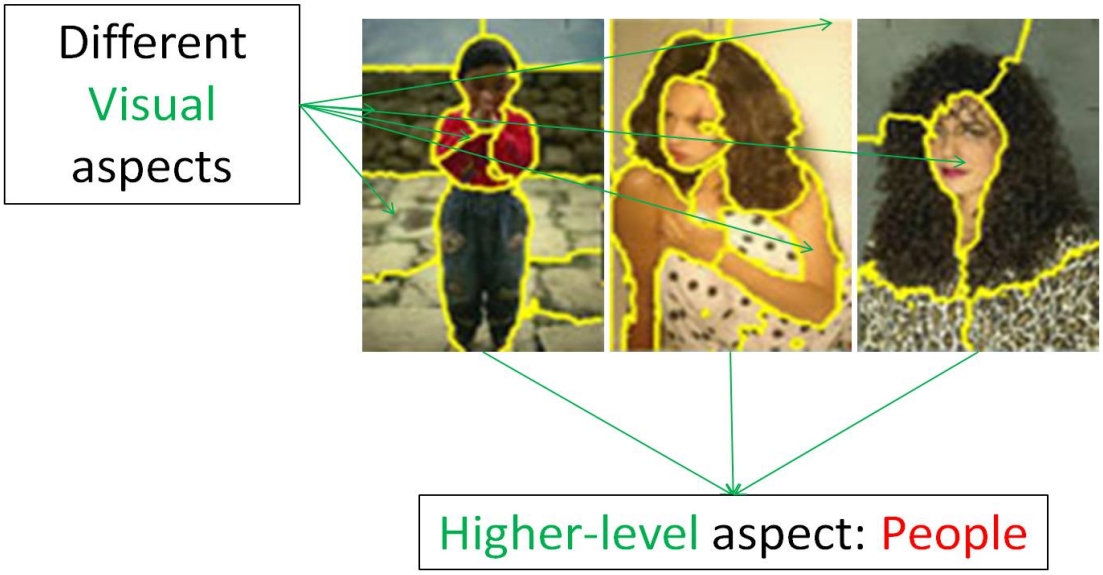

Figure 6.1: Examples of different visual and higher-level aspects.

In the MSSA model, there are two layers of latent topics (the high latent topics and the visual latent topics). One layer represents the high aspects (i.e., image categories) and the other one represents the visual aspects (i.e., objects, parts of objects or scenes). Furthermore, this model is in correspondence with the current belief in hierarchical recurrent cortex models of our brain [57].

Recently, Lienhart et al. [86]have introduced a hierarchal topic model (mm-pLSA). Even though the mm-pLSA model introduced a new multilayer inference model, it uses an EM algorithm to derive the different parameters, which costs a high computational power for parameters initialization and estimation. In addition, this approach did not introduce any criterion to estimate the number of

different latent variables. However, in MSSA the number of latent topic models is estimated in advance.

6.3 Generative process

Suppose that we have N images $\{im_j\}_{j=1}^{N}$ in which M visual representation units (visual words) $\{VRU_i\}_{i=1}^{M}$ are observed.

We introduce the high latent topics and visual latent topics in the following generative process for a given image im_j:

- Choose a high latent topic h_k from $P(h_k|im_j)$, a multinomial distribution conditioned on im_j and parameterized by a $K \times N$ stochastic matrix θ, where $\theta_{kj} = P(h_k = k|im_j = j)$.
- Choose a visual latent topic v_l from $P(v_l|h_k)$, a multinomial distribution conditioned on h_k and parameterized by an $L \times K$ stochastic matrix φ, where $\varphi_{lk} = P(v_l = l|h_k = k)$.
- Generate a visual representation unit VRU_i from $P(VRU_i|v_l)$, a multinomial distribution conditioned on v_l and parameterized by an $M \times L$ stochastic matrix Ψ, where $\Psi_{il} = P(VRU_i = i|v_l = l)$.

This generative process leads to the following conditional probability distribution:

$$P(VRU_i|im_j) = \sum_{k=1}^{K}\sum_{l=1}^{L} P(h_k|im_j, \theta)P(v_l|h_k, \varphi)P(VRU_i|v_l, \Psi) \tag{6.1}$$

Following the maximum likelihood principle, one can estimate the parameters by

maximizing the log-likelihood function as follows:

$$Li = \sum_{j=1}^{N} \sum_{i=1}^{M} n(VRU_i, im_j) log(P(VRU_i|im_j)) \tag{6.2}$$

Where $n(VRU_i, im_j)$ denotes the number of the occurance of the VRU_i in im_j. Figure 6.2 depicts the generative process using the plate notation.

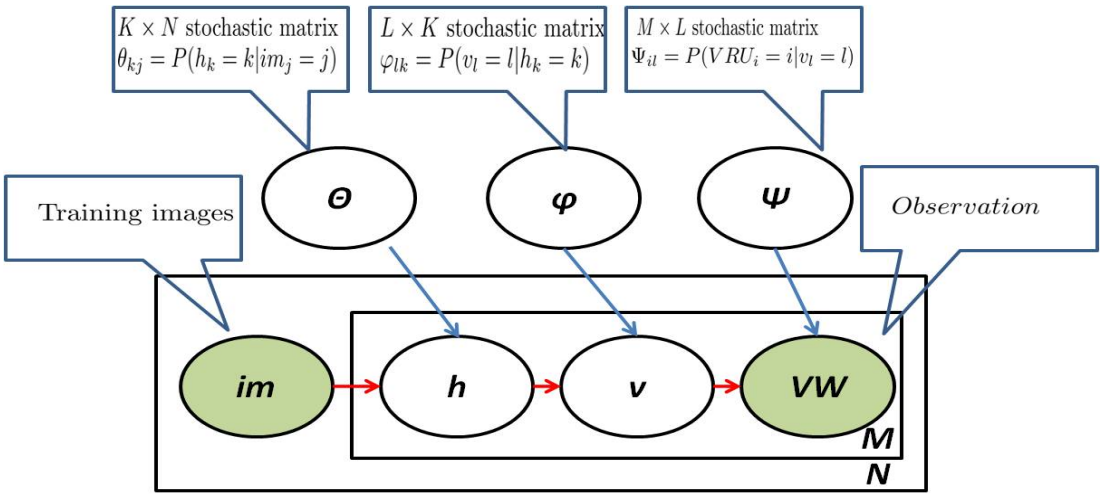

Figure 6.2: The semantic model using the plate notation.

6.4 Parameter estimation

The expectation-maximization (EM) algorithm [35] is the standard approach for maximum likelihood estimation in latent variable models. The main difficulty when implementing the EM algorithm in this work is that a four dimensional matrix is required in the E-step because of the two latent variables, which induces a high complexity.

However, Gaussier et al. [50] have proven that maximizing the likelihood

can be seen as a Non-Negative Matrix Factorization (NMF) problem under the generalized KL divergence. This leads to the following objective function:

$$min_{\theta,\varphi,\Psi} GL(A, \Psi\varphi\theta) \tag{6.3}$$

Where Ψ, φ, and θ are stationary points, A is the observation matrix, and $GL(A, \Psi\varphi\theta)$ is generalized KL divergence such that:

$$\theta \in \mathbb{R}_{+}^{K\times N}, \theta^T 1 = 1 \tag{6.4}$$

$$\varphi \in \mathbb{R}_{+}^{L\times K}, \varphi^T 1 = 1 \tag{6.5}$$

$$\Psi \in \mathbb{R}_{+}^{M\times L}, \Psi^T 1 = 1 \tag{6.6}$$

$$A_{ij} = \frac{n(VRU_i, im_j)}{\sum_{i,j} n(VRU_i, im_j)} \tag{6.7}$$

$$GL(A, \Psi\varphi\theta) = \sum_{i=1}^{M}\sum_{j=1}^{N}(A_{ij}\log\frac{A_{ij}}{[\Psi\varphi\theta]_{i,j}} - A_{ij} + [\Psi\varphi\theta]_{ij}). \tag{6.8}$$

6.4.1 Karush Kuhn Tucker (KKT) conditions

We use the Kuhn-Tucker (KKT) conditions [80] to derive the multiplicative update rules for minimizing (5.5) since it can be formulated as a constrained minimization problem with the following inequality constraints:

$$\Psi_{il} > 0 \tag{6.9}$$

$$\varphi_{lk} > 0 \tag{6.10}$$

$$\theta_{kj} > 0 \tag{6.11}$$

The necessary KKT conditions for a minimum of the constrained problem stated above are obtained by using the Lagrange multiplier method. Let α_{il}, β_{lk}, γ_{kj} be the Lagrangian multipliers associated with the constraints $\Psi_{il}, \varphi_{lk}, \theta_{kj}$ respectively. The KKT conditions require the following optimality conditions:

$$\frac{\partial GL(A, \Psi\varphi\theta)}{\partial \Psi_{il}} = \alpha_{il} \tag{6.12}$$

$$\frac{\partial GL(A, \Psi\varphi\theta)}{\partial \varphi_{lk}} = \beta_{lk} \tag{6.13}$$

$$\frac{\partial GL(A, \Psi\varphi\theta)}{\partial \theta_{kj}} = \gamma_{kj} \tag{6.14}$$

where:

$$\frac{\partial GL(A, \Psi\varphi\theta)}{\partial \Psi_{il}} = \sum_{j=1}^{N} \left\{ [\varphi\theta]_{lj} - \frac{A_{ij}}{[\Psi\varphi\theta]_{ij}} [\varphi\theta]_{lj} \right\} \tag{6.15}$$

$$\frac{\partial GL(A, \Psi\varphi\theta)}{\partial \varphi_{lk}} = \sum_{i=1}^{M} \sum_{j=1}^{N} \left\{ \Psi_{il}\theta_{kj} - \frac{A_{ij}}{[\Psi\varphi\theta]_{ij}} \Psi_{il}\theta_{kj} \right\} \tag{6.16}$$

$$\frac{\partial GL(A, \Psi\varphi\theta)}{\partial \theta_{kj}} = \sum_{i=1}^{M} \left\{ [\Psi\varphi]_{ik} - \frac{A_{ij}}{[\Psi\varphi\theta]_{ij}} [\Psi\varphi]_{ik} \right\} \tag{6.17}$$

This leads to the following:

$$\sum_{j=1}^{N} \left\{ [\varphi\theta]_{lj} - \frac{A_{ij}}{[\Psi\varphi\theta]_{ij}} [\varphi\theta]_{lj} \right\} = \alpha_{il} \tag{6.18}$$

$$\sum_{i=1}^{M} \sum_{j=1}^{N} \left\{ \Psi_{il}\theta_{kj} - \frac{A_{ij}}{[\Psi\varphi\theta]_{ij}} \Psi_{il}\theta_{kj} \right\} = \beta_{lk} \tag{6.19}$$

$$\sum_{i=1}^{M} \left\{ [\Psi\varphi]_{ik} - \frac{A_{ij}}{[\Psi\varphi\theta]_{ij}} [\Psi\varphi]_{ik} \right\} = \gamma_{kj} \tag{6.20}$$

The following complementary slackness conditions are also required:

$$\alpha_{il}\Psi_{il} = 0 \tag{6.21}$$

$$\beta_{lk}\varphi_{lk} = 0 \tag{6.22}$$

$$\gamma_{kj}\theta_{kj} = 0 \tag{6.23}$$

6.4.2 New multiplicative update rules for NMF

The minimization of the objective function (6.3), should be done with non-negativity constraints as described in Section 6.4.1. A multiplicative updating is an efficient way in such case since it can easily preserve the non-negativity constraints at each iteration. The proposed multiplicative updating algorithms for NMF associated with the objective functions (6.3) are given as follows:

Multiplying both sides of (6.18), (6.19), and (6.20) by Ψ_{il}, φ_{lk}, and θ_{kj} respectively, leads to the following:

$$\left[\sum_{j=1}^{N}\left\{[\varphi\theta]_{lj} - \frac{A_{ij}}{[\Psi\varphi\theta]_{ij}}[\varphi\theta]_{lj}\right\}\right]\Psi_{il} = \alpha_{il}\Psi_{ij} \tag{6.24}$$

$$\left[\sum_{i=1}^{M}\sum_{j=1}^{N}\left\{\Psi_{il}\theta_{kj} - \frac{A_{ij}}{[\Psi\varphi\theta]_{ij}}\Psi_{il}\theta_{kj}\right\}\right]\varphi_{lk} = \beta_{lk}\varphi_{ij} \tag{6.25}$$

$$\left[\sum_{i=1}^{M}\left\{[\Psi\varphi]_{ik} - \frac{A_{ij}}{[\Psi\varphi\theta]_{ij}}[\Psi\varphi]_{ik}\right\}\right]\theta_{kj} = \gamma_{kj}\theta_{ij} \tag{6.26}$$

Incorporating (6.24), (6.25), and (6.26) with (6.21), (6.22), and (6.23), leads to the following:

$$\left[\sum_{j=1}^{N}\left\{[\varphi\theta]_{lj}-\frac{A_{ij}}{[\Psi\varphi\theta]_{ij}}[\varphi\theta]_{lj}\right\}\right]\Psi_{il}=0 \tag{6.27}$$

$$\left[\sum_{i=1}^{M}\sum_{j=1}^{N}\left\{\Psi_{il}\theta_{kj}-\frac{A_{ij}}{[\Psi\varphi\theta]_{ij}}\Psi_{il}\theta_{kj}\right\}\right]\varphi_{lk}=0 \tag{6.28}$$

$$\left[\sum_{i=1}^{M}\left\{[\Psi\varphi]_{ik}-\frac{A_{ij}}{[\Psi\varphi\theta]_{ij}}[\Psi\varphi]_{ik}\right\}\right]\theta_{kj}=0 \tag{6.29}$$

This suggests the following iterative multiplicative update rules:

$$\Psi_{il}\leftarrow\Psi_{il}\frac{\sum_{j=1}^{N}\frac{A_{ij}}{[\Psi\varphi\theta]_{ij}}[\varphi\theta]_{lj}}{\sum_{j=1}^{N}[\varphi\theta]_{lj}} \tag{6.30}$$

$$\varphi_{lk}\leftarrow\varphi_{lk}\frac{\sum_{i=1}^{M}\sum_{j=1}^{N}\frac{A_{ij}}{[\Psi\varphi\theta]_{ij}}}{\sum_{i=1}^{M}\sum_{j=1}^{N}\Psi_{il}\theta_{kj}} \tag{6.31}$$

$$\theta_{kj}\leftarrow\theta_{kj}\frac{\sum_{i=1}^{M}\frac{A_{ij}}{[\Psi\varphi\theta]_{ij}}}{\sum_{i=1}^{M}[\Psi\varphi]_{ik}} \tag{6.32}$$

A small positive parameter ϵ, with value 10^{-9}, is added to (6.30), (6.31), and (6.32) in order to avoid division by zero as follows.

$$\Psi_{il}\leftarrow\Psi_{il}\frac{\sum_{j=1}^{N}\frac{A_{ij}}{[\Psi\varphi\theta]_{ij}}[\varphi\theta]_{lj}}{\sum_{j=1}^{N}[\varphi\theta]_{lj}+\epsilon} \tag{6.33}$$

$$\varphi_{lk}\leftarrow\varphi_{lk}\frac{\sum_{i=1}^{M}\sum_{j=1}^{N}\frac{A_{ij}}{[\Psi\varphi\theta]_{ij}}}{\sum_{i=1}^{M}\sum_{j=1}^{N}\Psi_{il}\theta_{kj}+\epsilon} \tag{6.34}$$

$$\theta_{kj}\leftarrow\theta_{kj}\frac{\sum_{i=1}^{M}\frac{A_{ij}}{[\Psi\varphi\theta]_{ij}}}{\sum_{i=1}^{M}[\Psi\varphi]_{ik}+\epsilon} \tag{6.35}$$

Also, some normalizing coefficients (λ, μ, and ν) are added to (6.33), (6.34),

and (6.35) with the aim of satisfying the normalization constraints:

$$\Psi_{il} \leftarrow \lambda \Psi_{il} \frac{\sum_{j=1}^{N} \frac{A_{ij}}{[\Psi\varphi\theta]_{ij}} [\varphi\theta]_{lj}}{\sum_{j=1}^{N} [\varphi\theta]_{lj} + \epsilon} \tag{6.36}$$

$$\varphi_{lk} \leftarrow \mu \varphi_{lk} \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} \frac{A_{ij}}{[\Psi\varphi\theta]_{ij}}}{\sum_{i=1}^{M} \sum_{j=1}^{N} \Psi_{il} \theta_{kj} + \epsilon} \tag{6.37}$$

$$\theta_{kj} \leftarrow \nu \theta_{kj} \frac{\sum_{i=1}^{M} \frac{A_{ij}}{[\Psi\varphi\theta]_{ij}}}{\sum_{i=1}^{M} [\Psi\varphi]_{ik} + \epsilon} \tag{6.38}$$

The application of the final multiplicative update rules (6.36, 6.37, 6.38) find at least locally optimal solutions for the objective function (6.3), where all the different parameters (Ψ, φ, θ) are estimated.

Therefore, the semantic inferences of the observed visual representation units (visual words) are known and can be used for further semantic analysis. We would like to highlight that in this form, the proposed multiplicative update rules themselves are extremely easy to implement computationally.

6.5 Number of latent topics estimation

In the practical situations, the number of latent topics of given probabilistic topic model is usually not known in advance. To determine the number of latent topics of a model, the simplest approach is to examine the resulting divergence between the data and model. This methodology is based on the assumption that the divergence becomes smaller when the appropriate number of latent variables is selected. However, a model with a larger number of latent variables usually over-fits to the data, resulting a smaller divergence than the divergence with the

correct number of clusters.

In order to prevent this over fitting and to select the correct number of latent topics, it is necessary to add some value to the divergence values for models with a larger number of latent topics. This kind of penalization is not a straightforward task because the resulting optimal numbers of latent topics can vary with the amount of penalization.

In the proposed MSSA model, the likelihood of the fitted model can be calculated, so the well-established statistical theory of model selection can be applied directly. In this case, the amount of penalization can be determined based on the statistical theory. The Akaike information criterion (AIC) [4] and the Bayesian information criterion (BIC) [131] are model selection methods based on the statistical theory. We use both model selection methods for the MSSA model and their results are compared in Chapter 8. These criteria are relatively simple to apply because they require only the maximum likelihood achievable for a given model, rather than the likelihood throughout the parameter space. Of course, such simplification comes at a cost, the cost being that they are derived using various assumptions, particularly the gaussianity (or near-gaussianity) assumption of the posterior distribution, which may be poorly respected in real-world situations.

We also study another method based on the statistically theory; the Minimum Description Length (MDL) principle is an alternative to estimate the latent variables number. MDL is similar to BIC; however, it provides a natural safeguard against over fitting, because it implements a tradeoff between the complexity of the hypothesis (model class) and the complexity of the data given the hypothesis. The three criteria are described in details in the following Sections.

6.5.1 Akaike Information Criterion (AIC)

The Akaike information criterion is a measure of the relative goodness of fit of a statistical model. It was developed by Hirotsugu Akaike [4]. It is grounded in the concept of information entropy by offering a relative measure of the information loss when a given model is used to describe the reality. The main idea of AIC is to select the model that minimizes the negative likelihood penalized by the number of parameters as specified in the following equation:

$$AIC = -2Li + 2m_k \tag{6.39}$$

In MSSA, Li is the log-likelihood function expressed in (6.2), and m_k is the number of the free parameters needed, expressed as follows:

$$m_k = ML + LK + KN \tag{6.40}$$

Where M is the visual vocabulary size, L is the number of the visual latent topics, K is the number of high latent topics, and N is the number of the images in the dataset.

Given a data set, several candidate MSSA models with different numbers of latent topics (different values of K and L) are ranked according to their AIC values. The preferred MSSA model is the one with *the minimum AIC value*. Hence, the AIC does not only reward goodness of fit, but also includes a penalty that is an increasing function of the number of estimated parameters. This penalty discourages over fitting (increasing the number of free parameters in the model improves the goodness of fit, regardless of the number of free parameters in the

data-generating process).

When observing the AIC values, one may infer that the top two candidate models are roughly in a tie and the rest are far worse. Thus, AIC provides a means for comparison among model candidates with different number of parameters. However, AIC does not provide a test of a model in the usual sense of testing a null hypothesis; in other words, AIC can not estimate how well a model fits the data in an absolute sense. For example, with AIC would not be possible to detect a situation where all candidate models fir poorly.

6.5.2 Bayesian Information Criterion (BIC)

In statistics, the Bayesian information criterion (BIC) [131] or Schwarz criterion (also SBC, SBIC) is a criterion for model selection among a class of parametric models with different numbers of parameters. Choosing a model to optimize BIC is a form of regularization.

When estimating model parameters using the maximum likelihood estimation, it is possible to increase the likelihood by adding parameters, which may result in over fitting as discussed before in Section 6.5.1. In the same way as AIC, the BIC resolves this problem by introducing a penalty term for the number of parameters in the model as follows:

$$BIC = -2Li + m_k \ln T \tag{6.41}$$

Here again, Li is the log-likelihood function expressed in (6.2), m_k is the number of the free parameters needed as in (6.40), and T is the number of data points used in the parameter estimation, expressed as follows:

$$T = \sum_{j=1}^{N} \sum_{i=1}^{M} n(VRU_i, im_j) \tag{6.42}$$

Where $n(VRU_i, im_j)$ denotes the number of the times the observed visual representation units occur in im_j, N denotes the number of the images in the dataset, and M is the visual vocabulary size. The BIC assumes that the data points are independent and identically distributed which may not be valid depending on the dataset under consideration.

BIC is closely related to the Akaike information criterion (AIC), with the difference that the penalty term is larger in the BIC than in the AIC.

The AIC and the BIC do have the same aim of identifying good models even if they differ in their exact definition of a *good model*. Comparing them is thus justified, at least to examine how each criterion performs for the recovery of the correct model, or to study how they behave when both agree on selecting the same model.

6.5.3 Minimum Description Length (MDL) principle

The Minimum Description Length (MDL) [124] principle is a relatively recent method for an inductive inference that provides a generic solution to the model selection problem. MDL is based on the following insight: any regularity in the data can be used to compress the data, i.e. to describe it using fewer symbols than the number of symbols needed to describe the data literally. The more regularities there are, the more the data can be compressed. Equating learning with *finding regularity*, we can therefore say that the more we are able to compress the data, the more we have learned about the data.

MDL is very strongly connected to probability theory and statistics through the correspondence between codes and probability distributions. The goal of a statistical inference may be cast as trying to find *regularity* in the data. *Regularity* may be identified with *ability to compress*. MDL combines these two insights by viewing learning as data compression: it states that, for a given set of models H and a data set D, we the best model in H is the one that yields a maximum compression for D that compresses D most.

The Minimum Description Length (MDL) principle is expressed as the following:

$$MDL = Li - \frac{m_k}{\log(NM)} \tag{6.43}$$

Once again, the first term is the log-likelihood function expressed in (6.2), m_k is the number of the free parameters needed and defined at in (6.40), N is the number of the images in the dataset, and M is the visual vocabulary size.

Given a data set D, several MSSA candidate models with different number of latent topics are ranked according to their MDL values. The preferred MSSA model is the one with *the minimum MDL value*. Because of this principle, when different MSSA candidate models with different K and L values fit the data equally well, the simpler model is selected.

The main difference between model criteria, such as AIC and BIC, and the supervised discretization methods such as MDL that the supervised discretization methods provide a natural safeguard against over fitting, because the winning model is the one with the lowest combined complexity or description length [75]. In Chapter 8, we compare the results of AIC, BIC, and MDL in selecting the best

MSSA model for different datasets.

6.6 Summary and conclusion

In BOW models for images, the vocabulary creation process, based on clustering algorithms such as K-Means, is quite coarse and can lead to many insignificant visual words. Such words add ambiguity in the representation of the objects and the scenes, and then reduce the effectiveness of classification or retrieval processes.

In this chapter, the new Multilayer Semantically Significant (MSSA) Model is introduced in order to study the semantic inference of different atomic visual representation units (e.g. visual words) in order to select the semantically significant units from the visual vocabulary.

First, we discuss the motivation of proposing a new multilayer probabilistic model, MSSA, based on two different latent topic layers (the high latent topics and visual latent topics). Second, we define the different generative processes that model the probabilistic distribution of different elements in the MSSA model. Third, we employ the KKT conditions to derive new multiplicative update rules in order to estimate the parameters of the MSSA model. Finally, the number of latent topics in the MSSA model is estimated using different mode selection criterions: Akaike information criterion (AIC), Bayesian information criterion (BIC), and Minimum Description Length (MDL). The performance of the three criteria are compared in Chapter 8.

Chapter 7

Semantically Significant Invariant Visual Glossary (SSIVG) Representation

Contents

7.1 Overview

In BOW representation, the feature quantization process can lead to many ambiguous and insignificant visual words. Another evident drawback is that a given visual word might represent different semantic meanings in different image contexts. This encumbers the distinctiveness of visual words and leads to low discrimination power. In addition, the images of the same semantic class can have arbitrarily different visual appearances and shapes. Such visual diversity of object causes one image semantic to be represented by different visual words. This leads to low invariance of visual words.

Based on the BOW representation and the MSSA model introduced in Chapter 5 and Chapter 6 respectively, we address all the mentioned drawbacks by proposing a higher-level visual representation, the Semantically Significant Invariant Glossary (SSIVG) representation which is more discriminative and likely

invariant to the visual appearance difference. It also overcomes the feature quantization noisiness.

This chapter is organized as follows. In Section 7.2, we introduce the Semantically Significant Visual Words (SSVWs) that are selected from the constructed visual words in order to overcome the feature quantization noisiness. The selection process is based on the visual words semantic inferences that are estimated using the MSSA model. In Section 7.3, we strengthen the discrimination power of visual words by constructing Semantically Significant Visual Phrases (SSVPs) from frequently SSVW sets occurred in the same local context, involved in strong association rules, and semantically coherent. In section 7.4, we enhance the intra-class invariance power of the SSVWs and the SSVPs by clustering them based on their probability distributions to the relevant visual topics. This leads to form the Semantically Significant Invariant Visual Glossary (SSIVG) representation. In Section 7.5, we establish an original spatial weighting scheme that is associated with the vector space image model for image retrieval and indexing using SSIVG representation. We introduce a new Multilayer Vote-Based Classifier (MVBC) based on the SSIVG representation in Section 7.6. We give a summary and a conclusion of this chapter in Section 7.7.

7.2 Semantically Significant Visual Words (SSVWs) generation

The SSVWs are selected from set of visual words (VWs), as described below.

7.2.1 Selecting the SSVWs

After generating the VWs as described in Chapter 5, the MSSA model is run with the co-occurrence matrix of the visual words as the observation matrix A. This leads to estimate the different probability distributions $P(h_k|im_j)$, $P(v_l|h_k)$, and $P(VW_i|v_l)$. Subsequently, all the visual latent topics v_l are categorized according to their conditional probabilities with all the high latent topics $P(v_l|h_k)$. All the visual latent topics whose conditional probabilities relating to all the high latent topics are higher than a given threshold $t_{h_k}$are categorized as relevant. Given a set of the relevant visual topics, a Semantically Significant Visual Word (SSVW) is defined as follows.

Definition 6 (Semantically Significant Visual Word (SSVW))**.** *An SSVW is a visual word (VW) whose conditional probability $P(VW_i|v_l)$ is higher than a given threshold t_{v_l} for at least one relevant visual latent topic.*

From our perspective, all the visual words whose probability distributions $P(VW_i|v_l)$ are low for every relevant visual topic are irrelevant, since they are not informative for any relevant visual topic. Hence, we propose to keep only the most significant visual words for each relevant visual topic.

7.2.2 Examples of the SSVWs

Figure 7.1 gives examples of images displayed with VWs and SSVWs. The images in the left sides are images displayed with VWs. On the right side, the same images are displayed SSVWs. The huge difference in the number of VWs and the number of SSVWs is obvious since about 30% of the VWs are selected

as SSVWs. It is clear that most of the SSVW are describing different part of the main objects (dog and dinosaur).

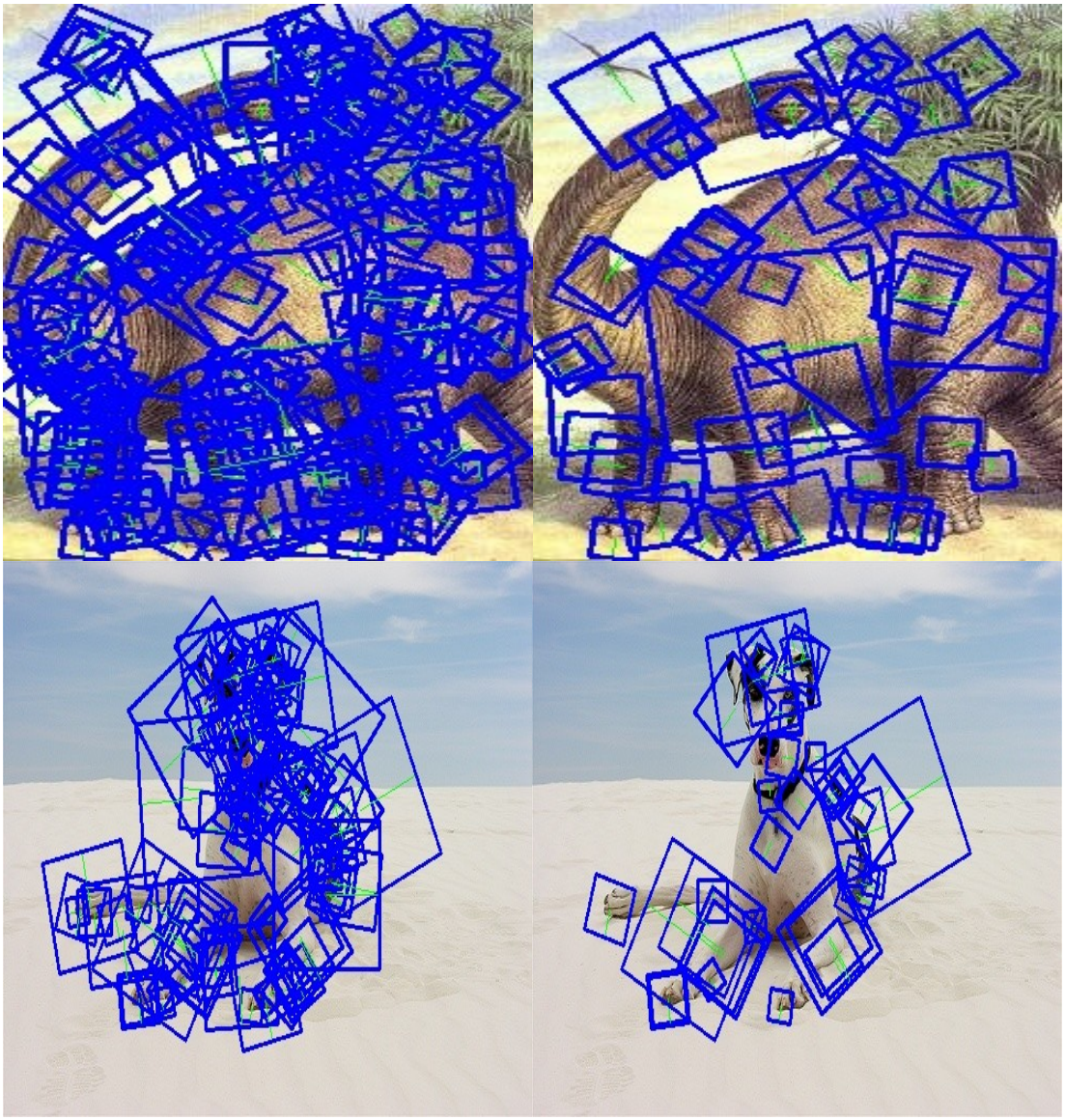

Figure 7.1: The left side of the figure is an example of two images displayed with all constructed VWs and the right side is the same images displayed with SSVWs.

7.3 Semantically Significant Visual Phrases (SSVPs) generation

The low discrimination power of the SSVWs leads to low correlations between the image features and their semantics. Such low correlation motivates to generate a higher-level discriminative visual representation, named Semantically Significant Visual Phrase (SSVP). Analogous to text documents, which are particular arrangements of words in 1D space, images can be seen as particular arrangements of patches in a 2D space.

There are theoretical similarities between natural languages and visual languages. A natural language consists of words, and a visual language consists of visual words. In natural languages, there are grammars, which restrict the words distribution and order. In an image, divided into patches, there exists some constraints about how the patches are combined together to form meaningful objects. Indeed, a random combination of patches or pixels does not construct a meaningful image. The SSVWs and their inter-relationships are the basis for generating SSVPs [38], which are defined as follows.

Definition 7 (Semantically Significant Visual Phrase (SSVP))**.** *We define an SSVP as a set of Semantically Significant Visual Words (SSVWs) that frequently co-occurr together in a spatial local context, involved in strong association rules, and semantically coherent.*

Since it is not easy to define the semantic coherence in a set of SSVWs, we assume the following:

Assumption 1 (Semantically Coherent Set of SSVWs)**.** *A set of SSVWs are se-*

mantically coherent whenever they have a high probability regarding to at least one common relevant visual latent topic. Their probability distributions are estimated using the MSSA model.

In this section, we discuss the different processes for generating the SSVPs. These processes start by defining the local neighborhood of a given SSVW, and finish by generating a representation scheme for the constructed SSVP vocabulary.

7.3.1 Low discrimination power of the SSVWs

An SSVW represents different semantic meanings in different image context. This encumbers the distinctiveness of the SSVWs and leads to a low discrimination. In fact, the discrimination issue is a problem of under-representation [172]. Its consequence is effectively small interclass distances [177]. One of the major reasons for the low discrimination issue is that the regions represented in a visual word might come from the object with different semantics but similar local appearance.

This can be explained as a polyseme problem by an analogy between text documents and visual documents. A polyseme is a word with multiple, related meanings and senses. For instance, box is defined as *financial institution*, *ground bounding waters*, or *row or tier of objects*. If polysemous words like this are considered in a text document representation, they can exert a deleterious effect for classification and retrieval accuracy because their ambiguity makes them have strong relationships with other unwanted categories by their different senses from the intended one. In other words, every text word should be fundamentally selected so as to characterize one category by its single meaning. If it has several

senses; the ambiguity causes the features to also characterize often non relevant categories.

Figure 7.2 gives an example of two SSVWs that share visually similarities in two different categories (*car* and *motorbike*). The SSVW A is, therefore, not able to distinguish *motorbike* from *car*. However, SSVW A and SSVW B considered

Figure 7.2: An example of the low discrimination power of the SSVWs.

together can effectively distinguish *motorbike* from *car*. The discrimination of representation can therefore be improved by mining interrelations among SSVWs in a certain neighborhood region in order to construct a more discriminative higher-level representation.

7.3.2 Spatial local context

Several methods have been proposed to sample spatial neighborhoods within an image. In [33], a sliding-window mechanism samples windows at a fixed location and scale step, followed by a spatial tiling of the windows. The very different approach [138] defines a neighborhood around each region. This is represented as

an unordered set of the K-nearest regions, without storing any spatial information (K-neighborhoods). However, the neighborhoods in this case are always of a fixed size.

Our approach attempts to combine the best of them. Instead of using a K-neighborhood, we use the *scale* of the center of the local patch to define the size of the neighborhood and all SSVWs (not just *pairs* of SSVWs) that occur within this context are considered in the SSVP generation process. Figure 7.3 shows examples of the local contexts around the center of different patches. The square represents a local patch; the red circle around the center of the local patch denotes the local context.

7.3.3 Frequent SSVW sets mining

Frequent sets play an essential role in many data mining tasks that try to find relevant patterns from databases, such as association rules, correlations, sequences, episodes, classifiers and clusters. The identification of sets of items, products, symptoms and characteristics, which often occur together in the given database, can be seen as one of the most basic tasks in data mining.

Historically, the original motivation for searching frequent sets came from the need to analyze so-called supermarket transaction data, that is, to examine customer behavior in terms of the purchased products [1]. Frequent sets of products describe how often items are purchased together.

As mentioned in Section 7.3, we define an SSVP as a set of SSVWs that frequently co-occurr together in a spatial local context, involved in strong association rules, and semantically coherent. Therefore, in this section we discuss

the different steps for discovering frequent SSVW sets that are involved in strong association rules.

7.3.3.1 Frequent SSVW sets discovery

To find the frequent SSVW sets within an image database, we consider the followings:

- I is a set of SSVWs that occur in the same spatial context.
- A transaction over I is a couple $T = (t_{id}, I)$ where t_{id} is the transaction identifier and I is the set of SSVWs.
- A database D over I is a set of transactions over I such that each transaction has a unique identifier.

A transaction $T = (t_{id}, I)$ is said to support a set X, if $X \subset I$. The *support* of a set X in D is the probability that X occurs in a transaction, or in other words, is the number of transactions that support X in D divided by the total number of transactions in the database D.

Definition 8 (Frequent SSVW Set)**.** *An SSVW set is called frequent SSVW set if its support is greater than a given minimal support threshold, min_supp, with 0<min_supp<1.*

The task of discovering all frequent sets is challenging. The search space is exponential with respect to the number of SSVWs occurring in the database, and the targeted databases can be massive, containing millions of transactions. Although a number of algorithms have been proposed to discover frequent item sets, Apriori algorithm [2] remains the most efficient [56], and therefore we select.

Apriori algorithm is a seminal algorithm, which uses an iterative approach known as a level-wise search, where k-itemsets are used to explore (k+1)-itemsets. It uses the Apriori property to reduce the search space: all non-empty subsets of a frequent itemset must also be frequent.

7.3.3.2 Association rules generating from frequent SSVW sets

Once the frequent SSVW sets from transactions in a database D have been found, it is straightforward to generate strong association rules from them, where a strong association rule is defined as a rule whose support and confidence satisfy *minimum support* (min_supp) and *minimum confidence* (min_conf) respectively [101]. The confidence and support of a rule can be estimated as follows:

$$support(A \Rightarrow B) = P(A \cup B) = \frac{support\_count(A \cup B)}{|D|} \tag{7.1}$$

$$confidence(A \Rightarrow B) = P(B|A) = \frac{support\_count(A \cup B)}{support\_count(A)} \tag{7.2}$$

The conditional probability is expressed in terms of frequent SSVW sets *support_count*, where:

- $support\_count(A \cup B)$ is the number of transactions containing the frequent SSVW sets $A \cup B$.
- $support\_count(A)$ is the number of transactions containing the frequent SSVW set A.

We generate the association rules from the frequent SSVW set as follows:

- For each frequent SSVW sets I, generate all nonempty subsets.
- For each non-empty subset s of I, output the rule $s \Rightarrow (I - s)$ if

$\frac{support\_count(I)}{support\_count(s)} >= min\_conf$,

Knowing that $support(s \Rightarrow (I-s)) = support(I)$ since $s \cup (I-s) = I$. Therefore, $(s \Rightarrow (I-s)) = support(I) >= min\_supp$ since I is a frequent SSVW set.

After generating the strong association rules from the frequent SSVW sets, the semantic coherence of the SSVWs that are involved in strong association rules is checked within each set. As mentioned before, we assume that a set of SSVWs are *semantically coherent whenever they have a high probability relating to at least one common relevant visual latent topic*. Their probability distributions are estimated using the MSSA model.

Finally, the generated SSVP set is composed of SSVW sets that satisfy all the following conditions:

- they are frequently occur in the same local context.
- they are involved in strong association rules.
- they have a high probability relating to the same visual latent topic.

7.3.4 Examples of the SSVPs

Figure 7.3 shows examples of SSVPs corresponding to three different visual aspects. Here again, the square represents a local patch; the red circle around the center of the local patch denotes the local context, and the group of local patches in the same context denotes an SSVP.

7.3.5 SSVP vocabulary construction

For the purpose of online indexing and retrieval, we need an efficient representation scheme to describe and store the SSVP vocabulary. We design a simple

Figure 7.3: Examples of SSVPs appearing in different images.

but efficient method based on *hashing*. A hash map that contains the indexes of all SSVPs is constructed to map groups of frequent SSVW sets (that are involved in strong association rules, semantically coherent, and are within the same

local context in a given image) to visual phrases. The key is the *base 36* of c, where c is the concatenation of the constituent visual words indexes after sorting. Figure 7.4 represents an example of five SSVWs $SSVW_{2065}$, $SSVW_{621}$, $SSVW_{1191}$, $SSVW_{2130}$, $SSVW_{775}$ mapped to $SSVP_{122}$ that has a hash key = $4Q28VUFALILE$ (base 36 of 621775119120652130). This internal representation scheme offers us several important benefits.

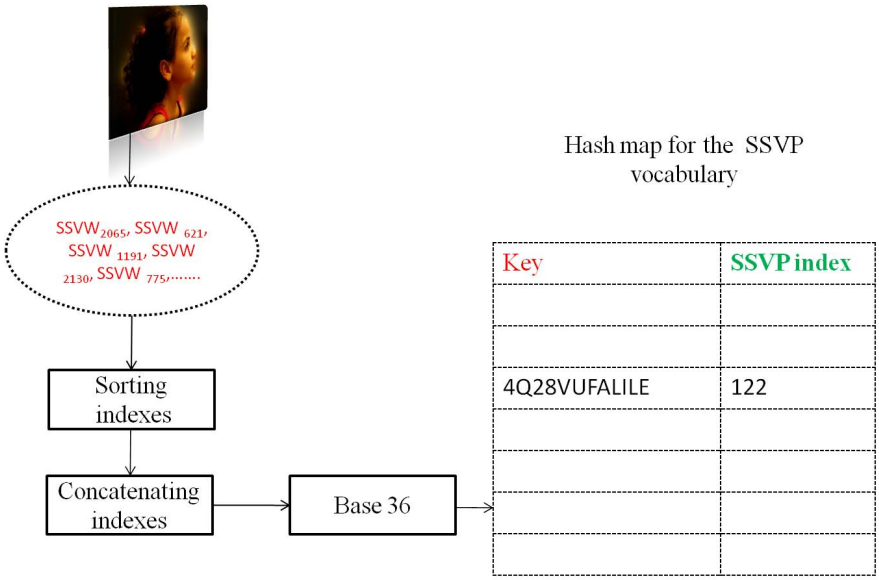

Figure 7.4: An example of five SSVWs mapped to an SSVP.

Firstly, the hash mapping of the SSVWs to SSVPs is much faster than using a plain list of the SSVPs, and it is also better from the binary search. For instance, binary search can locate an item in a sorted table of n items with $log_2 n$ key comparisons. Therefore, this hash map will be more efficient than binary search since no comparison with other items is needed.

Secondly, the choice of 36 is convenient and compact in that the digits can be represented using the Arabic numerals $0-9$ and the Latin letters $A-Z$. Thus, we less memory allocation and the algorithm is consequently more efficient.

7.4 Semantically Significant Invariant Visual Glossaries (SSIVGs) generation

Even though studying the co-occurrence and spatial scatter information makes the image representation more distinctive, the invariance power of SSVWs or SSVPs is still low. Returning to text documents, synonymous words are usually clustered into one synonym set to improve the document categorization performance [12]. Such an approach inspires us to partially bridge the visual diversity of the images by clustering the SSVWs and the SSVPs based on their probability distributions to all relevant visual latent topics.

After the distributional clustering, *each group of SSVWs that belongs to a given cluster are re-indexed with the same index as the cluster centroid*. This leads to generate Semantically Significant Invariant Visual Words (SSIVWs) which consist of SSVWs that are re-indexed after distributional clustering. In the same manner we generate the Semantically Significant Invariant Visual Phrase (SSIVP). Finally, both the SSIVWs and the SSIVPs form the Semantically Significant Invariant Visual Glossary (SSIVG) representation.

Definition 9 (Semantically Significant Invariant Visual Glossary (SSIVG) representation)**.** *Semantically Significant Invariant Visual Glossary (SSIVG) representation is a higher-level visual representation composed from two different layers of representation: Semantically Significant Invariant Visual Word (SSIVW) representation and Semantically Significant Invariant Visual Phrase (SSIVP) representation, where an SSIVW (resp. SSIVP) is an SSVW (resp. SSVP) that has been re-indexed after a distributional clustering.*

In this section, we discuss the invariance problem. Then, the MSSA model is run one more with the new observations that are built from the co-occurrence of the SSVWs and the SSVPs in order to estimate the new probability distributions for both of them. Based on the estimated probabilistic inferences, we cluster SSVs and the SSVPs. Finally, the SSVWs and SSVPs are re-indexed to form the SSIVW and SSIVP respectively.

7.4.1 Low invariance of the SSVWs and SSVPs

The images in a given semantic class can have arbitrarily different visual appearances and shapes. Such visual diversity of objects causes one visual aspect to be possibly represented by different SSVWs and SSVPs. This leads to low invariance of SSVWs and SSVPS. The consequence is large intra-class variations. In this circumstance, the SSVS and SSVPS become too primitive to effectively model the image semantics, as their efficacy depends highly on the visual similarity and regularity of images of the same semantics.

Figure 7.5 gives an example of the invariance problem in two images of motorcycles. The two different SSVWs ($SSVW_{607}$, $SSVW_{1076}$) occurring in the two images describe the same part of the of motorcycle, however they are different, and therefore they have different indexes (607 and 1076). Also, the two SSVPs ($SSVP_{148}$, $SSVP_{263}$) are describing the same part of the motorcycle (part of the wheels), and they are different indexes (148 and 263). This happens since the two images are for the same object (motorcycle), yet with different shapes and colors. This leads to extract different low-level features from the two images. In text domain, when documents of a same topic or categories contain different sets

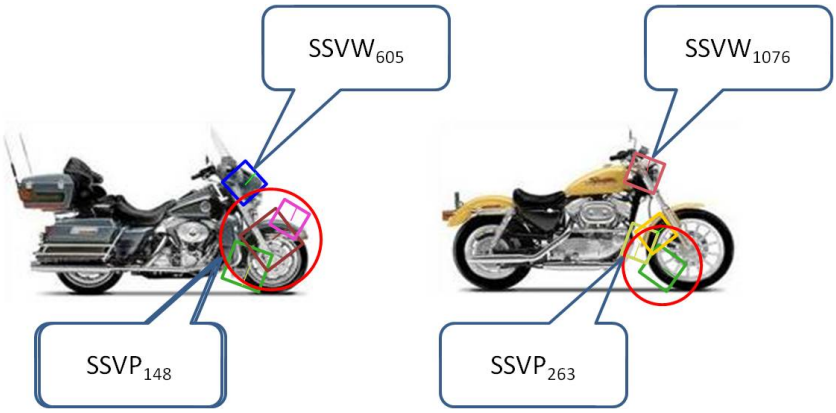

Figure 7.5: Illustration of the invarince problem: similar image regions are indexed with different SSVWs and SSVPs

of words, the word synset (synonym set) that links words of similar semantics is robust to model them [12]. Inspired by this, we propose that relevance-consistent group of the SSVWs or SSVPs with similar semantic inferences should have the same index.

7.4.2 New generative process

After generating the SSVWs and the SSVPs, the co-occurrence of both forms new observations. We study the semantic inferences for the SSVWs and the SSVPs after the new observations. The same MSSA model that is introduced in Chapter 6 is run with the co-occurrences of the SSVWs and the SSVWs as the observation matrix.

After running the MSSA according to the above generative processes, the new probability distributions for SSVWs and SSVPs to different visual latent topics are estimated.

7.4.3 Distributional clustering for SSVWs and SSVPs

After estimating the new semantic inferences of the SSVWs and the SSVPs, the next step is to group the SSVWs that are with similar probabilistic inferences. Similarly, the SSVPs that share alike semantic inferences are also grouped. In our approach, we use an information-theoretic framework that was introduced by Dhillon et al. [36]. This framework is similar to Information Bottleneck [32] by deriving a global criterion, that captures the optimality of distributional clustering. The main criterion is based on the generalized Jensen-Shannon divergence [87] among multiple probability distributions.

Let v_l be a discrete random variable that takes on values from the set of relevant visual latent topics V=$\{v_1,v_2,...,v_L\}$. Let SSVW_m and $\text{SSVP}_{m'}$ be the random variables that range over the set of the SSVWs $\{\text{SSVW}_1,\text{SSVW}_2,...,\text{SSVW}_M\}$ and the set of SSVPs $\{\text{SSVP}_1,\text{SSVP}_2,...,\text{SSVP}_{M'}\}$ respectively.

Since we are interested in reducing the number of features and the model size, we only consider the *hard* clustering where each SSVW (resp. SSVP) belongs to exactly one SSVW cluster (reps. SSVP). We cluster the SSVWs into K clusters according to their probability distributions. In the same manner, we cluster the SSVPs into Q clusters as follows.

Let the random variable c_k range over the SSVW clusters $C = \{c_1, c_2, ..., c_K\}$. To judge the quality of word clusters an information-theoretic measure is used. The information about $SSVW_m$ captured by v_l can be measured by the mutual information $I(SSVW_m; v_l)$. Ideally, in forming SSVW clusters, we would like to exactly preserve the mutual information; however, a non-trivial clustering always lowers mutual information. Dhillon et al. proposed to find a clustering that

minimizes the decrease in mutual information, $I(SSVW_m; v_l) - I(c_k; v_l)$, for a given number of SSVW clusters.

Dhillon et al. prove that the decrease in mutual information can be expressed in terms of the generalized Jensen-Shannon divergence of each cluster as follows:

$$I(SSVW_m; v_l) - I(c_k; v_l) = \sum_{k=1}^{K} \pi(c_k) JS_{\pi'}(\{P(v_l/SSVW_t) : SSVW_t \in c_k\}) \tag{7.3}$$

Where:

$$\pi(c_k) = \sum_{SSVW_t \in c_k} \pi(SSVW_t) \tag{7.4}$$

$$\pi(SSVW_t) = P(SSVW_t) \tag{7.5}$$

$$\pi_t^{'} = \frac{\pi_t}{\pi(c_k)} \tag{7.6}$$

and JS denotes the generalized Jensen-Shannon divergence.

The generalized Jensen-Shannon divergence of a finite set of probability distributions can be expressed as the (weighted) sum of Kullback-Leibler divergences to the (weighted) mean, as follows:

$$JS_{\pi}(\{p_i : 1 \leq n \leq i\}) = \sum_{i=1}^{n} \pi_i KL(p_i, m) \tag{7.7}$$

Where $\pi_i \geq 0$, $\sum_i \pi_i = 1$ and m is the weighted mean probability distribution, $m = \sum_i \pi_i p_i$.

By (7.3) and (7.7), the decrease in mutual information due to word clustering

can be written as follows:

$$\sum_{k=1}^{K} \pi(c_k) \sum_{SSVW_t \in c_k} \frac{\pi_t}{\pi(c_k)} KL(p(SSVW_i|v_l)(p(c_k|v_l)) \tag{7.8}$$

Where the probability distribution $p(c_k|v_l)$ is estimated as follows:

$$p(c_k|v_l) = \sum_{SSVW_t \in c_k} \frac{\pi_t}{\pi(c_k)} p(SSVW_t|v_l) \tag{7.9}$$

As a result, the quality of SSVW clustering can be measured by the following objective function:

$$Q(\{c_k\}_{k=1}^{K}) = I(SSVW_i; v_l) - I(c_k; v_l) = \sum_{k=1}^{K} \sum_{SSVW_t \in c_k} \pi_t KL(p(SSVW_t|v_l))(p(c_k|v_l)) \tag{7.10}$$

According to Dhillon et al., writing the objective function in the above manner suggests an iterative algorithm that repeatedly does the following:

- Re-partition the distributions $p(SSVW_t|v_l)$ by their closeness in KL- divergence to the cluster distributions $p(c_k|v_l)$.
- Subsequently, given the new word clusters, re-computes these cluster distributions using (7.8).

Algorithm (1) describes the Divisive Information Theoretic Clustering algorithm in details, as it is used in our approach. Dhillon et al. showed that their algorithm minimizes *within-cluster divergence* and simultaneously maximizes *between-cluster divergence*. This approach is markedly better than the agglomerative algorithm of Baker and McCallum [8] and the one introduced by Slonim and Tishby [139].

Algorithm 1 Divisive Information Theoretic Clustering (P, ψ, l, k, W)

Input:
P is the set of distributions, $p(SSVW_t|v_l) : 1 \leq t \leq M$,
Π is the set of all SSVW priors, $\pi = p(SSVW_t) : 1 \leq t \leq M$,
L is the number of visual latent topics,
K is the number of desired clusters.
Output:
C is the set of word clusters $c_1, c_2, ..., c_K$.

1. Initialization: for every SSVW $SSVW_t$, assign $SSVW_t$ to C_q such that $p(SSVW_t|v_l) = maxip(SSVW_i|v_l)$.This gives L' initial SSVW clusters; if $Q \geq L$ split each cluster arbitrarily into at least $\lfloor K/L \rfloor$ clusters, otherwise merge the L' clusters to get Q SSVW clusters.
2. For each cluster c_k, compute $\pi(c_k) = \sum_{g_t \in c_k} \pi(SSVW_t)$, $p(c_k|v_l) = \sum_{SSVW_t \in c_k} \frac{\pi_t}{\pi(c_k)} p(SSVW_t|v_l)$.
3. Re-compute all clusters: For each $SSVW_t$, find its new cluster index as $j * (SSVW_t) = argmin_i KL(p(SSVW_t|v_l), p(c_k|v_l))$, resolving ties arbitrarily.
   Thus compute the new SSVW clusters $c_k, 1 \leq k \leq K$, as
   $c_k = SSVW_t : j * (SSVW_t) = k$.
4. Stop if the change in objective function value given by (7.10) is *small* (10^{-3});
   Else go to step 2.

We cluster the SSVPs to Q clusters in the same manner using the same Divisive Information Theoretic Clustering algorithm (1) stated above.

7.4.4 Semantically Significant Invariant Visual Words and Phrases (SSIVWs and SSIVPs) generation

After the distributional clustering, each group of SSVWs that tends to share similar probability distributions are grouped in the same cluster c_k and re-indexed with the same index k. In the same manner, each group of SSVPs that share similar probability distributions are clustered in the same cluster c_q and re-indexed

with same index q.

After re-indexing the SSVWs and the SSVPs, they form the Semantically Significant Invariant Visual Words (SSIVWs) and the Semantically Significant Invariant Visual Phrases (SSIVPs) respectively. Both of the SSIVW and the SSIVPs form the Semantically Significant Visual Glossaries (SSIVGs)

By generating the SSIVG representation, the visual differences of images from the same class can be *partially* bridged. Consequently, the image distribution in the feature space will become more coherent, regular and stable.

7.5 Image indexing and retrieval using the SSIVG representation

Inspired by the success of the vector-space model in the text document representation, it is applied recently to the image representation. As mentioned in Chapter 4, each image is represented by a k-dimensional vector of the estimated weights associated with the visual index terms appearing in the image collections. Many effective information retrieval weighting schemes are applied to vector-space model in order to estimate the weights of the visual index terms such as $tf \times idf$, weighting scheme.

Most of the weighting schemes do not integerate the spatial location of the local patches. However, the spatial aspects in an image, carry important information for classifying and retrieving the images [170, 70]. For example, an image showing a beach scene typically consists of sky-like local patches on the top, and sands-like local patches in the bottom. The $tf \times idf$ weighting scheme does not

take in consideration such spatial information and may result in inferior classification performance. In this section, we introduce a new spatial weighting scheme for the SSVWs in order to integrate the spatial information. This spatial weighting scheme is associated with the Vector Space Model that we use it for different levels of representations.

7.5.1 A new spatial weighting scheme for the SSIVWs

In order to allow a ***spatial weighting*** for the SSIVWs, we design a new scheme that is a variation of the $tf \times idf$ weighting scheme. Suppose that in an image, there are local features obtained from the interest point set belonging to a given Gaussian and assigned to an SSIVW$_l$, where 1<l<IVW and IVW is the SSIVWs vocabulary. The sum of the probabilities of salient point occurrences indicate the contribution of the SSIVW$_l$ to a Gaussian g_i. Therefore, the weighted term frequency ($Tf_{SSIVW_l g_i}$) of an SSIVW$_l$ with respect to the Gaussian g_i is defined as follows:

$$Tf_{SSIVW_l g_i} = \sum_{m=1}^{n_l} P(g_i|f_m) \tag{7.11}$$

Where n_l denotes the number of the occurrence of SSIVW$_l$ in a given Gaussian g_i, and f_m is the local feature that corresponds to an occurrence of SSIVW$_l$.

The average weighted term frequency (Tf_{SSIVW_l}) of SSIVW$_l$ with respect to an image I where $SSIVW_l$ occurs in n_{SSIVW_l} Gaussian is defined as follows:

$$Tf_{SSIVW_l} = \sum_{i=1}^{n_{SSIVW_l}} (Tf_{SSIVW_l g_i})/n_{SSIVW_l} \tag{7.12}$$

The weighted inverse Gaussian frequency of SSIVW$_l$ with respect to an image

I with n Gaussian is defined as follows:

$$If_{SSIVW_l} = ln\frac{n}{n_{SSIVW_l}} \tag{7.13}$$

The final spatial weight of the visual word $SSIVW_l$ is defined by the following formula:

$$Sw_{SSIVW_l} = Tf_{SSIVW_l} \times If_{SSIVW_l} \tag{7.14}$$

7.5.2 Vector space image model

The traditional Vector Space Model [129] of Information Retrieval [157] is adapted to our representation, and used for similarity matching and retrieval of images. The following doublet represents each image in the model:

$$I = \begin{cases} \overrightarrow{SSIVW_i} \\ \overrightarrow{SSIVP_i} \end{cases} \tag{7.15}$$

where $\overrightarrow{SSIVW_i}$ and $\overrightarrow{SSIVP_i}$ are the vectors for the word and phrase representations of a document respectively:

$$\overrightarrow{SSIVW_i} = (SSIVW_{1,i},\ ...,\ SSIVW_{n_SSIVW,i})$$

$$\overrightarrow{SSIVP_i} = (SSIVP_{1,i},\ ...,\ SSIVP_{n_SSIVP,i}) \tag{7.16}$$

Note that the vectors for each level of representation lie in a separate space. In the above vectors, each component represents the weight of the corresponding

dimension. We used the ***spatial weight scheme*** defined in Section 7.5.1, for the SSIVWs and the standard ***td×idf weighting scheme*** for the SSIVPs.

In our approach, we use an inverted file [164] to index images. The inverted index consists of two components: one includes the visual index terms (SSIVW and SSIVP), and the other includes vectors containing the information about the spatial weighting of the SSIVW and the $tf \times idf$ weighting of the SSIVP.

7.5.3 Similarity measure

The query image is represented as a doublet of SSIVWs and SSIVPs and we consult the inverted index to find candidate images. All candidate images are ranked according to their similarities to the query image. We have designed a simple measure that allows evaluating the contribution of words and phrases. The similarity measure between a query I_q and a candidate image I_c is estimated with:

$$sim(I_q, I_c) = (1-\alpha)RSV(\overrightarrow{SSIVW_c}, \overrightarrow{SSWIV_q}) + (\alpha)RSV(\overrightarrow{SSIVP_c}, \overrightarrow{SSIVP_q}) \tag{7.17}$$

The Retrieval Status Value (RSV) of 2 vectors is estimated with the cosine distance. The non-negative parameter $0 < \alpha < 1$ is to be set according the experiment runs in order to evaluate the contribution between the SSIVWs and the SSIVPs.

7.6 Multiclass Vote-Based Classifier (MVBC)

We propose a new multiclass vote-based classification technique (MVBC) based on the SSIVG representation. For each $SSIVG_i$ occurring in an image im_j, we detect the high latent topic h_k that maximizes the following conditional probability:

$$p(SSVG_i|h_k) = p(v_l|h_k)p(SSIVG_i|v_l) \tag{7.18}$$

The final voting score VS_{h_k} for a high latent topic h_k in a test image imj_j is :

$$VS_{h_k} = \sum_{a=1}^{N_{h_k}^{SSIVG}} p(SSIVG_a|h_k) \tag{7.19}$$

Where $N_{h_k}^{SSIVG}$ is the number of SSVGs voted for h_k in im_j. Finally, each image is categorized according to the dominant high latent topic which is the topic with the highest voting score (the high latent topic and the class labels are mapped in the training dataset).

7.7 Summary and conclusion

The standard BOW representation has led to many significant results in various vision tasks including object recognition and categorization. However, in practice, the clustering of low-level local features leads to some insignificant or noisy visual words. In addition, this representation suffers from low discrimination and invariance powers.

In this chapter, we tackle these draw backs by proposing a higher-level visual representation, the Semantically Significant Invariant Glossary (SSIVG) represen-

tation. This representation is based on the BOW representation and the MSSA model introduced in Chapters 5 and 6 respectively.

We introduce the Semantically Significant Visual Words (SSVWs) that are chosen from the constructed visual words in order to overcome the feature quantization nosiness. The selection process is based on the visual words semantic inferences that are estimated using the MSSA model. In addition, the discrimination power of the SSVWs is strengthened by building Semantically Significant Visual Phrases (SSVPs) from frequently co-occurring SSVW sets occurred in the same local context, involved in strong association rules, and semantically coherent. Moreover, we boost the intra-class invariance power of the SSVWs and the SSVPs by clustering them based on their probability distributions to the relevant visual topics. This leads to form the Semantically Significant Invariant Visual Glossary (SSIVG) representation. Besides, we establish a new spatial weighting scheme that is associated with the vector space image model for image retrieval and indexing using the SSIVG representation. Furthermore, we introduce a new Multilayer Vote-Based Classifier (MVBC) based on the SSIVG representation.

Part III

Experimental Results and Applications

Chapter 8

Experimental Results

Contents

8.1 Overview

We have implemented the proposed approach of a higher-level visual representation, and the system is evaluated. This chapter reports the large-scale, extensive experimental evaluations of the in comparison with the state-of-the-art literature to demonstrate the superiority of the proposed methods in the context of three different applications, image retrieval, classification, and object recognition.

This chapter is organized as follows. In Section 8.2, we introduce the three datasets NUS-WIDE dataset that are used in the context of image retrieval, classification, and object recognition respectively. Subsequently, we discuss the estimation of the different parameter settings. In Section 8.3, we extensively examine the performance of the proposed higher-level of representation for a image retrieval. We compare the performance of each different layer of the proposed representation (Enhanced-BOW, SSVW, SSVP, SSIVW, SSIVP, and SSIVG). We also extend the performance comparison to several other recently proposed higher-level representation methods in image retrieval context. In Section 8.4, we study the performance of the proposed SSIVG representation and the proposed MVBC classifier in the context of image classification. In Section 8.5, we evaluate the performance of the proposed combination of the SSIVG representation and the MVBC classifier in the object recognition task. Finally, we give a summary and conclusion for this chapter in Section 8.6.

8.2 Dataset and experimental setup

8.2.1 Datasets

We use three different datasets in our experiments for different applications as follows.

- We evaluate the proposed SSIVG representation on image retrieval using the NUS-WIDE dataset [29], one of the largest available datasets with 269,648 images and the associated tags from Flickr website. We separate the dataset into two parts. The first part contains 161,789 images to be used for training and the second part contains 107,859 images to be used for testing. It contains 81 image categories as shown in Figure 8.1.
- We have tested the proposed MVBC and the SSIVG representation on the MIRFLICKR-25000 [64] dataset for classification. The dataset contains 25000 images that were retrieved from the Flickr website. The images are annotated with 11 general topics. The general topics were chosen in such a way that they mostly correspond to common Flickr tags themselves and may contain some additional common tags as subtopics. The general topics and corresponding subtopics selected are listed in Table 8.1. We have used the 11 general annotations as ground truth for image classification. We use 15000 images as a training dataset from different image classes and the rest 10000 images for testing.
- Caltech101 dataset [45] is used the proposed SSIVG representation in object recognition. It contains 8707 images, which include objects belonging to 101 classes. Table 8.2 lists the 101 object category of the Caltech101 dataset. The number of images ranges from 40 to 800 images per category. Most

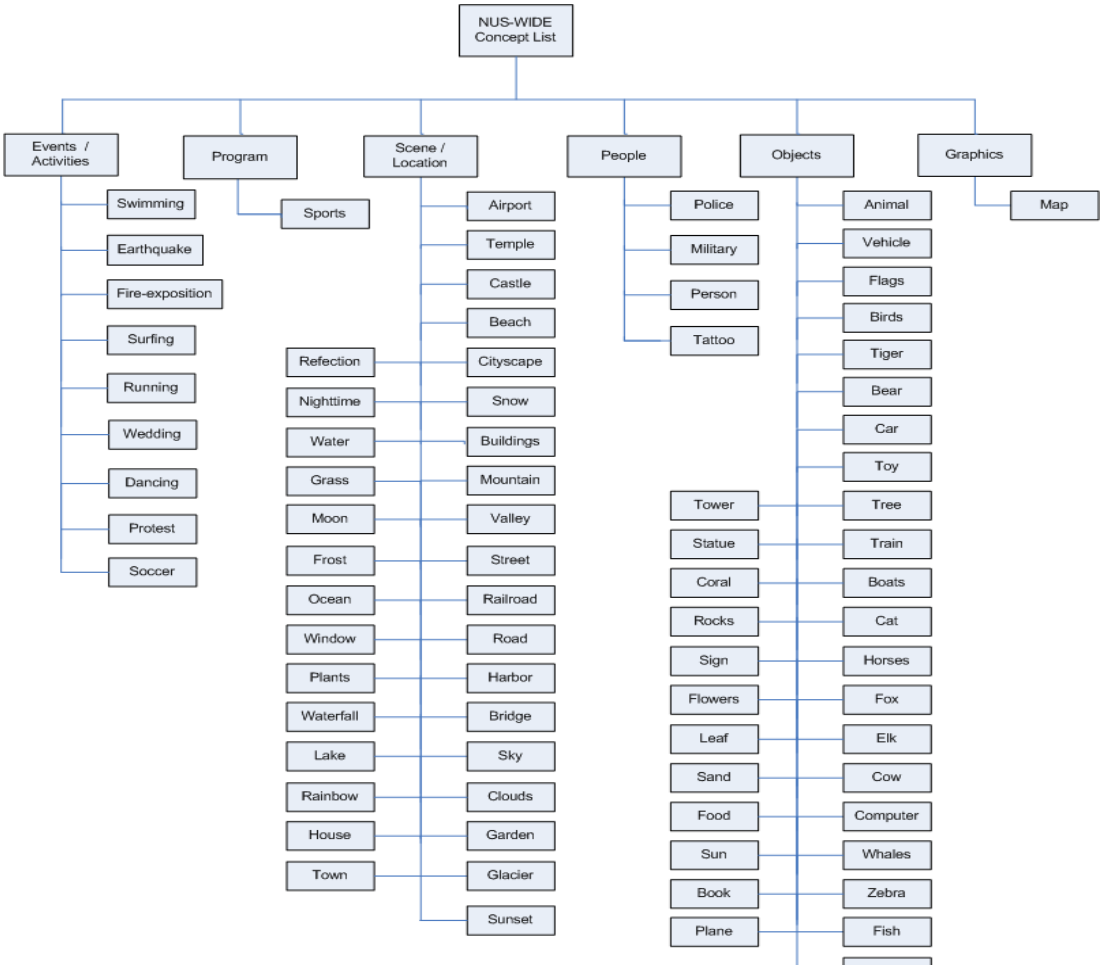

Figure 8.1: The concept taxonomy of NUS-WIDE.

categories have about 50 images. For the various experiments, we construct the test dataset by selecting randomly 10 images from each object category (resulting in 1010 images) and we select 30 images from each object category (different from the test images) for the training.

General topics	Subtopics
sky	clouds
water	sea/ocean, river, lake
people	portrait, boy/man, girl/woman, baby
night	
plant life	tree, flower
animals	dog, bird
man-built structures	architecture, building, house, city/urban, bridge, road/street
sunset	
indoor	
transport	car

Table 8.1: The general topics and corresponding subtopics selected in the MIRFLICKR-25000.

trilobite	face	pagoda	tick	inlineskate	metronome
accordion	yinyang	soccerball	spotted cat	nautilus	grand-piano
crayfish	headphone	hawksbill	ferry	cougar-face	bass
ketch	lobster	pyramid	rooster	laptop	waterlilly
wrench	strawberry	starfish	ceilingfan	seahorse	stapler
stop-sign	zebra	brontosaurus	emu	snoopy	okapi
schooner	binocular	motorbike	hedgehog	garfield	airplane
umbrella	panda	crocodile-head	llama	windsor-chair	car-side
pizza	minaret	dollarbill	gerenuk	sunflower	rhino
cougar-body	crab	ibis	helicopter	dalmatian	scorpion
revolver	beaver	saxophone	kangaroo	euphonium	flamingo
flamingo-head	elephant	cellphone	gramophone	bonsai	lotus
cannon	wheel-chair	dolphin	stegosaurus	brain	menorah
chandelier	camera ant	scissors	butterfly	wild cat	lamp
crocodile	barrel	joshua-tree	pigeon	watch	dragonfly
mayfly	cup	ewer	octopus	platypus	buddha
chair	anchor	mandolin	electric-guitar		

Table 8.2: 101 Caltech object category list.

8.2.2 Evaluation criteria

8.2.2.1 Image retrieval context

The evaluation criteria used in the image retrieval context is the mean average precision (MAP), which is the mean value of *average precision* (AP) of each query. The AP is the sum of the precision values at each relevant hit in the

retrieval list, divided by the total number of relevant images in the collection. AP_q is defined for a query q as:

$$AP_q = \frac{\sum_{r=1}^{R_q} Prec(r) \times rel(r)}{T_q} \tag{8.1}$$

Where r is image rank, R_q is the total number of images retrieved, $Prec(r)$ is the precision of retrieval list cut-off at rank r, $rel(r)$ is an indicator (0 or 1) of the relevance of rank r, and T_q is the total number of relevant images for q in the corpus. The average precision is an ideal measure of retrieval quality, which is determined by the overall ranking of relevant images. Intuitively, the MAP gives higher penalties to fault retrievals if they have higher position in the ranking list. This is rational, as in practice, searchers are more concerned with the retrieved results in the top.

8.2.2.2 Image classification and object recognition context

In the context of image classification tasks, the notions true positive, true negative, false positive and false negative are used to compare a given classification of an image (the class label assigned to the image by a classifier) with the desired correct classification (the class where the image actually belongs).

To measure classification performance, we used the classification *average precision* (AP) [99] over each image class. It is a popular measure that takes into account both recall and precision values since it is equivalent to the area under the precision-recall curve. It is computed as follows:

– A version of the measured precision-recall curve is estimated with the precision monotonically decreasing, by setting the precision for recall r to the

maximum precision obtained for any recall $r' \geq r$.

– The AP is computed as the area under this curve by numerical integration.

For an object recognition task, each test image is recognized by predicting the object class using the SSIVG representation and the MVBC. Thus, the same criteria as for image classification are used here.

8.2.3 Parameters estimation

By changing different parameters of the experimental setting, several aspects can be investigated which have influence on the performance of the proposed visual representation in the context of different applications. In this section, we discuss the different parameter settings of the different datasets.

8.2.3.1 Visual word vocabulary size

The generation of the proposed higher-level visual representation (SSIVG representation) is a bottom-up process. Hence, selecting the proper visual word vocabulary size at the lower level of representation (bag of visual words representation) is essential to the whole process. In this Section, we investigate the proper values of different visual word vocabulary sizes corresponding to the different datasets.

Unlike the vocabulary of a text corpus whose size is relatively fixed, the number of clusters in the local feature quantization process controls the size of a visual word vocabulary. Choosing the right vocabulary size involves the trade-off between the discriminative power and the computational cost. With a small vocabulary, many of the visual words are not discriminative because dissimilar local features can map to the same visual word. Using a large vocabulary increases the

cost of clustering local features, computing visual-word features, and running the MSSA model. Hence, there is no consensus for the appropriate size of a visual word vocabulary.

The visual word vocabulary size used in other works varies from several hundred [81], to thousands and tens of thousands [137]. Their results are not directly comparable due to the difference on corpus and classification or retrieval methods. To find out the proper visual word vocabulary sizes corresponding to the different datasets, we study the influence of the visual vocabulary size on the performance of the enhanced bag of visual words introduced in Chapter 5 within the context of the different applications.

Figure 8.2 shows the mean average precision in the context of image retrieval using the NUS-WIDE dataset for different values of the corresponding visual vocabulary size (K). The traditional Vector Space Model of Information Retrieval is adapted using the inverted file structure, and the $tf \times idf$ weighting for the constructed visual words. It is clear that the highest MAP is at K=10000. In addition, we can see that when K changes from 5000 to 10000, the MAP value for the system drastically increases from 0.165 to 0.193. This shows that the performance of the system is ***sensitive*** to visual word vocabulary size K.

Figure 8.3 shows the mean average precision in the context of image classification using MIRFLICKR-25000 dataset for different values of the corresponding visual vocabulary size $K^{'}$. We use the SVM with a linear kernel as a classifier and $tf \times idf$ as weighting scheme. The figure shows that the highest MAP is at $K^{'} = 3000$. Also, it is obvious that the performance of the BOW representation is so sensitive the K in the context of classification.

Figure 8.4 shows the mean average precision in the context of object recog-

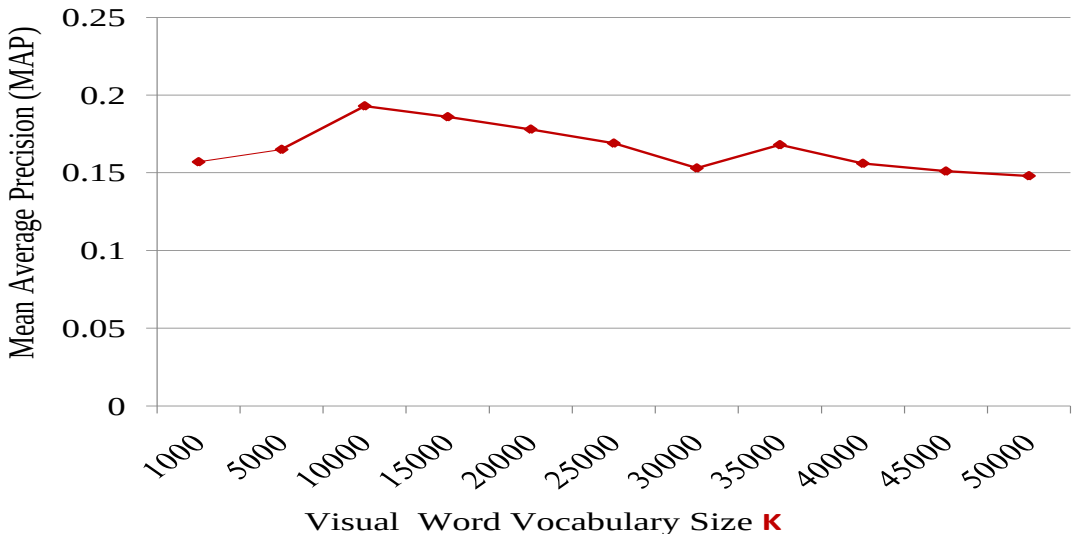

Figure 8.2: Evaluation for the visual vocabulary size for retrieval on NUS-WIDE dataset.

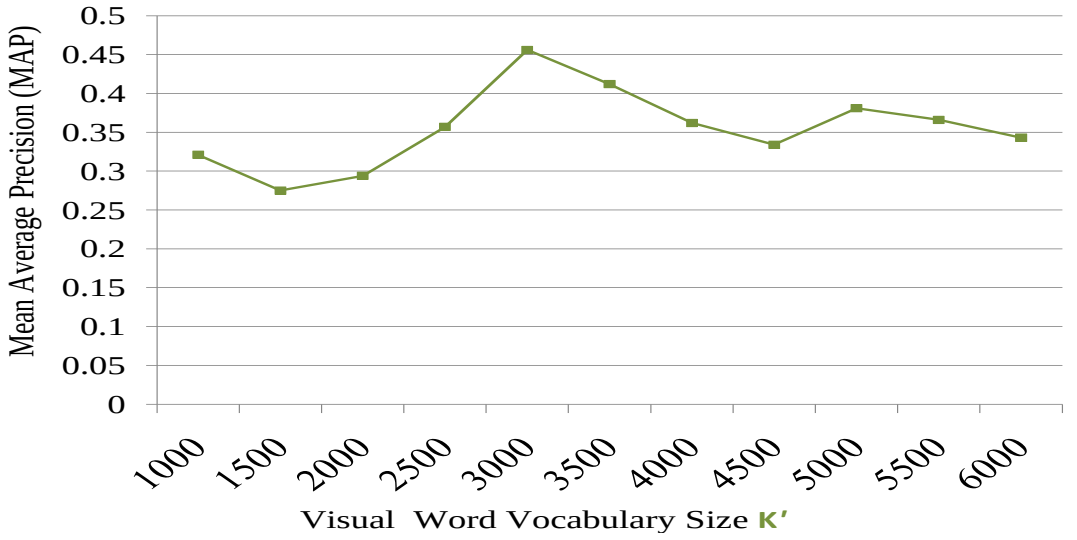

Figure 8.3: Evaluation for the visual vocabulary size for classification on MIRFLICKR-2500 dataset.

nition using Caltech101 dataset for different values of the corresponding values of the visual vocabulary size K″. We also use the SVM with a linear kernel as a classifier and $tf \times idf$ as weighting scheme to predict the object class for each test image. According to the results, the highest Map is at K″ = 2750. We can always see that when K″ changes, the MAP value for the system extremely changes. Hence, choosing the suitable visual word vocabulary is important since it affects the performance of the system in different context (image classification, image retrieval, and object recognition).

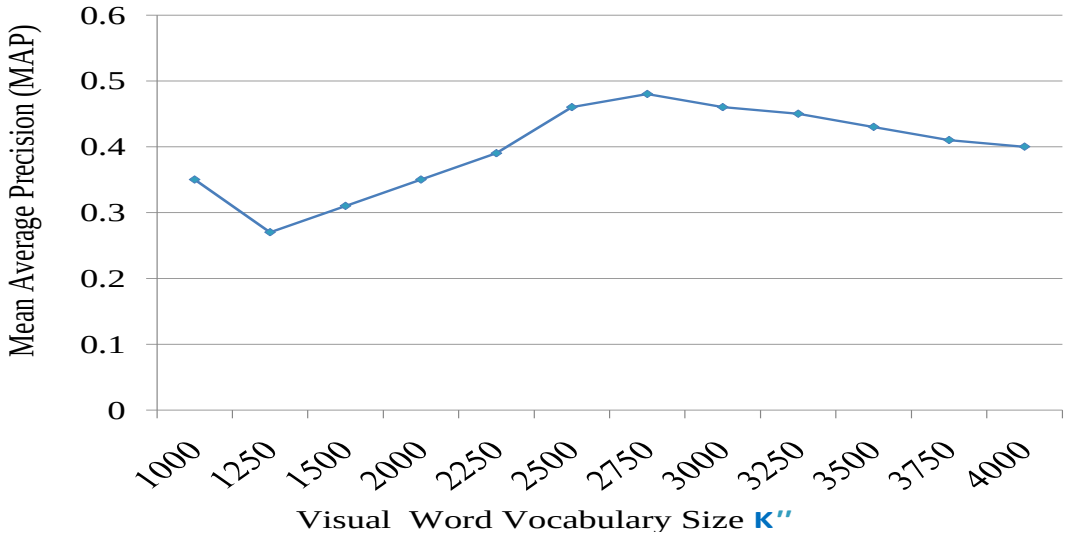

Figure 8.4: Evaluation for the visual vocabulary size for object recognition on Caltech101 dataset.

8.2.3.2 Number of latent topics

As mentioned in Chapter 6, we estimate the number of high and visual latent topics in the MSSA model are estimated using three different model selection techniques: Akaike information criterion (AIC), Bayesian information criterion

(BIC), and Minimum Description Length (MDL) Principle. In this section, we evaluate the performance of these techniques in estimating the number of high and visual latent topics using for the three datasets (NUS-WIDE, MIRFLICKR-25000, and Caltech101). In our approach, the high latent topics represent the class labels of the training dataset. The numbers of the class labels for the three datasets are known from the ground truth. Thus, we evaluate the performance of AIC, BIC and MDL based on the correspondence between the estimated numbers of high latent topics and the actual number of the class labels. We try different number of high and visual latent topics, and we compute AIC, BIC and MDL values for each case.

Figure 8.5, Figure 8.6, and Figure 8.7 show the AIC values for NUS-WIDE, Caltech101, MIRFLICKR-25000 datasets respectively. As a result, AIC find neither the true number of high latent topics nor close number since the number of the high latent topics corresponding to the maximum value of AIC for NUS-WIDE dataset, MIRFLICKR-25000 dataset, and Caltech101 dataset are 60, 80, and 80 respectively which are far from the numbers of class labels in the ground truth. As mentioned in Section 8.2.1, the number of the class labels of NUS-WIDE dataset, MIRFLICKR-25000 dataset, and Caltech101 dataset are 81, 11, and 101 respectively.

As shown in Figure 8.8, Figure 8.9, and Figure 8.10 BIC performs better than AIC using MIRFLICKR-2500 dataset (the estimated number of high latent topics is 10) but fails to predict the suitable numbers of the high latent topics in the NU-SWIDE dataset (the estimated number of high latent topics are 50) and Caltech101 dataset (the estimated number of high latent topics is 70).

However, the maximum values of MDL that are shown in Figures 8.11, 8.12,

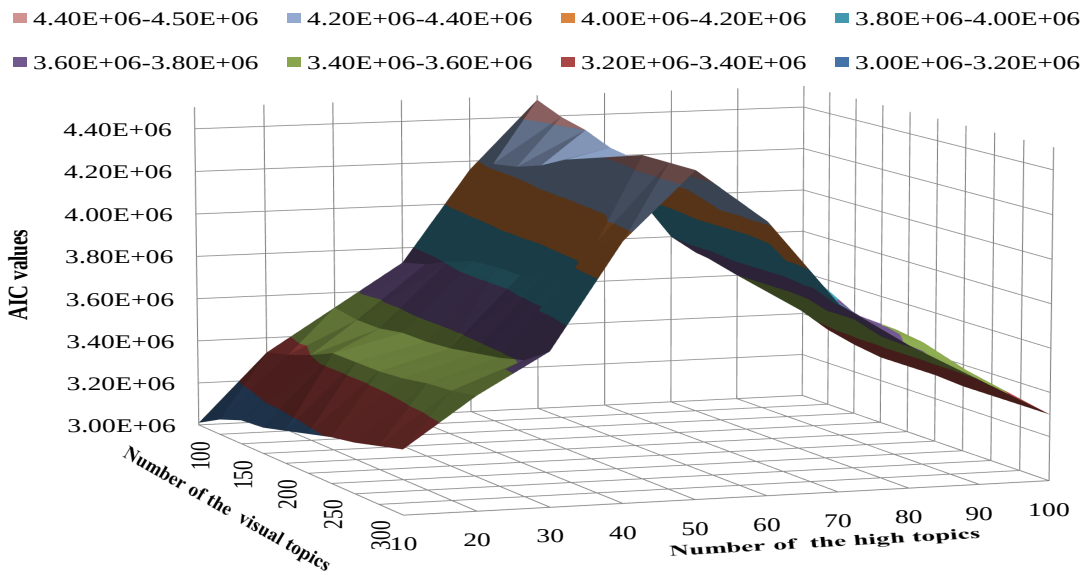

Figure 8.5: AIC values using the NUS-WIDE dataset.

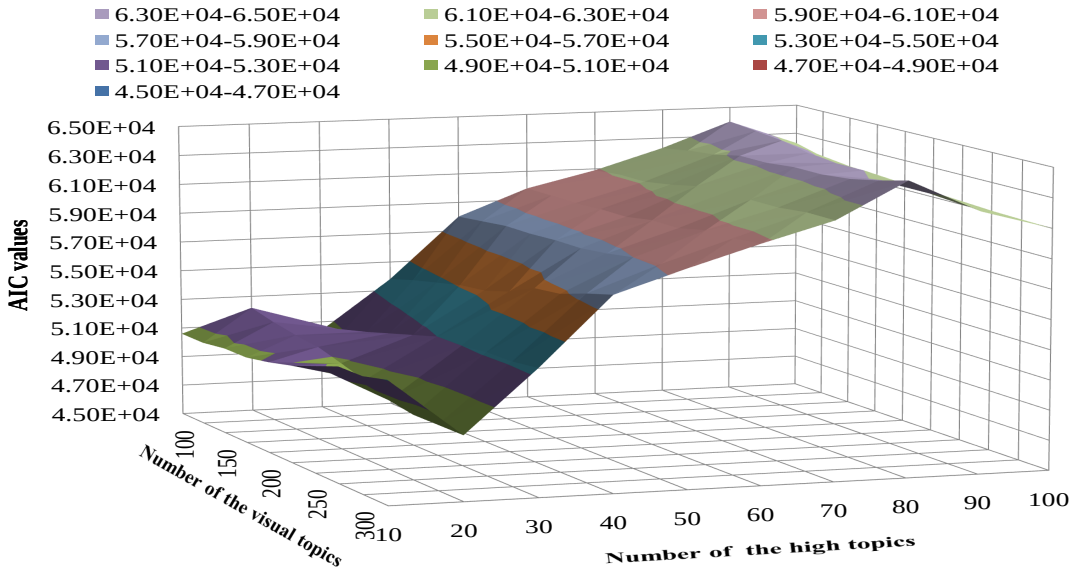

Figure 8.6: AIC values using the MIRFLICKR-25000 dataset.

and 8.13 clearly indicates a number of high latent topics that is close to the ground truth in the three datasets. Hence, we take the results of MDL as parameter

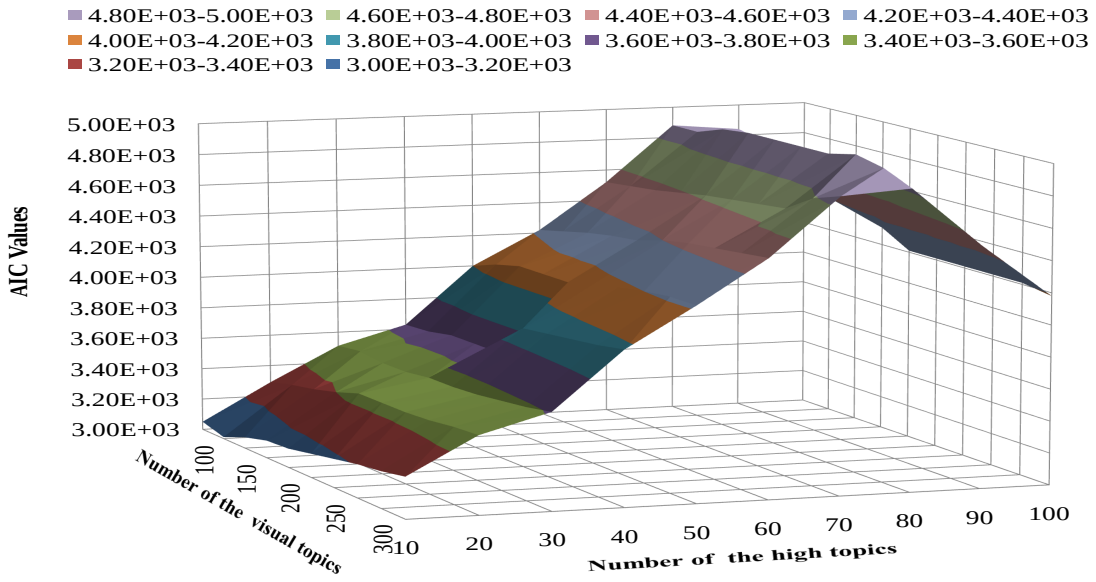

Figure 8.7: AIC values using the Caltech101.

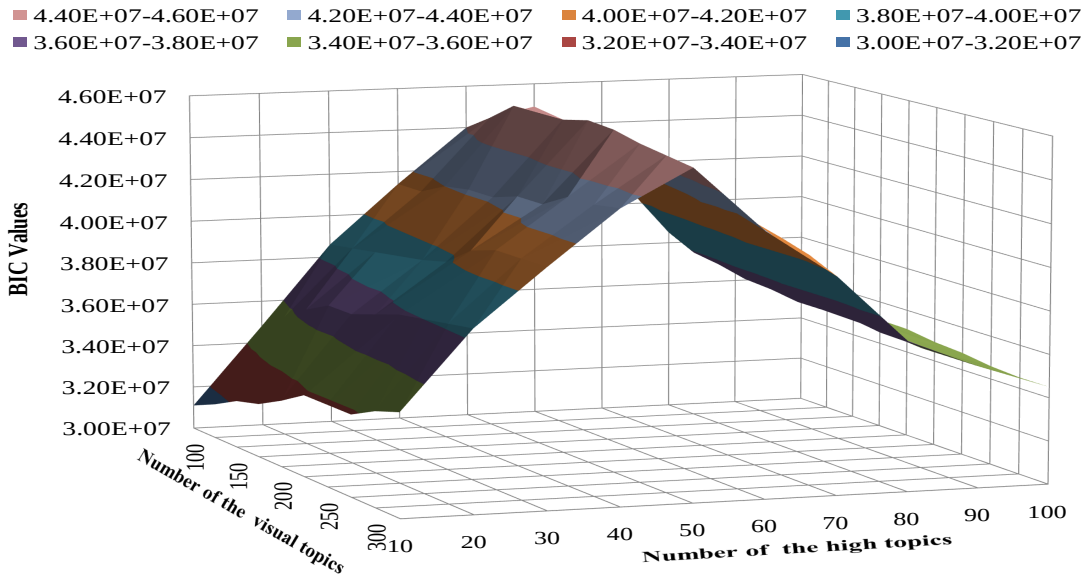

Figure 8.8: BIC values using the NUS-WIDE dataset.

settings of the MSSA model for the three datasets, as mentioned in Table 8.3.

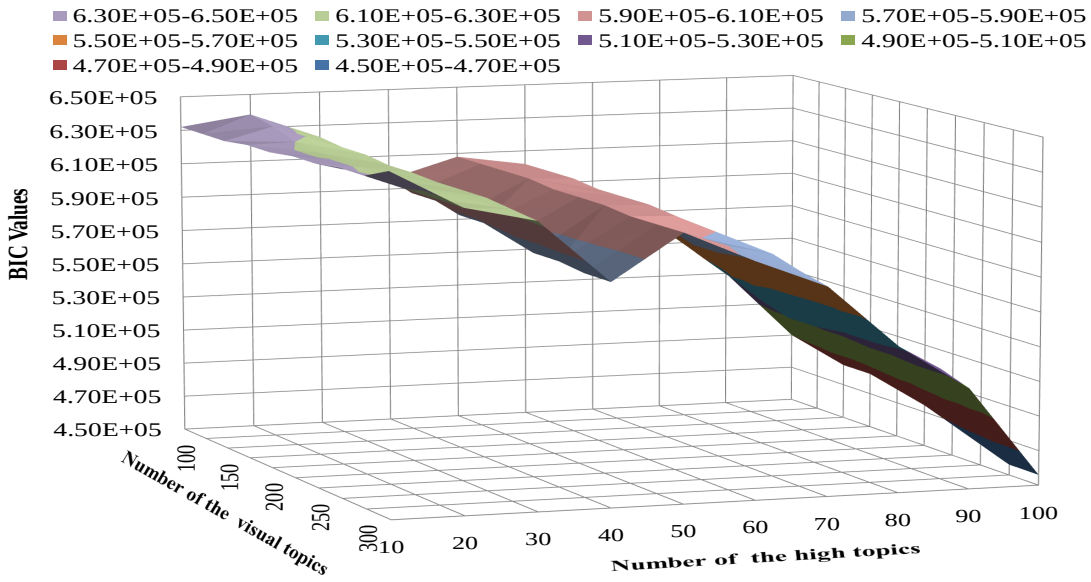

Figure 8.9: BIC values using the MIRFLICKR-25000 dataset.

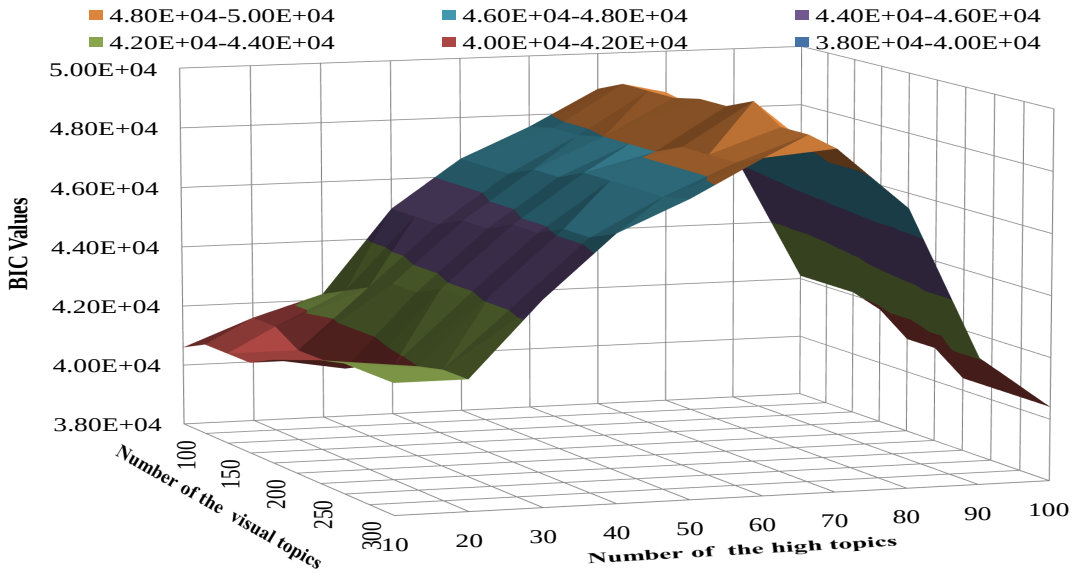

Figure 8.10: BIC values using the Caltech101.

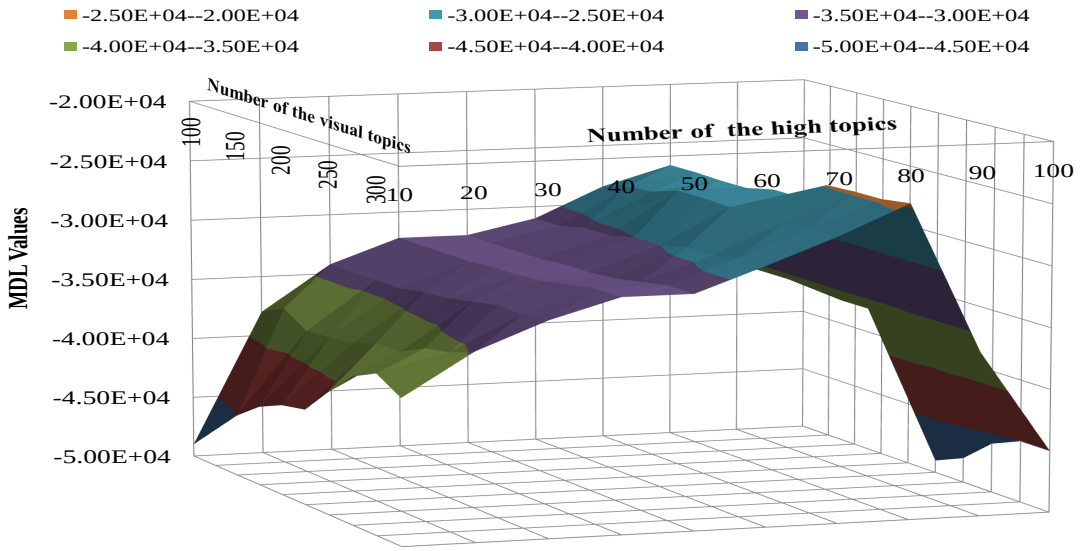

Figure 8.11: MDL values using the NUS-WIDE dataset.

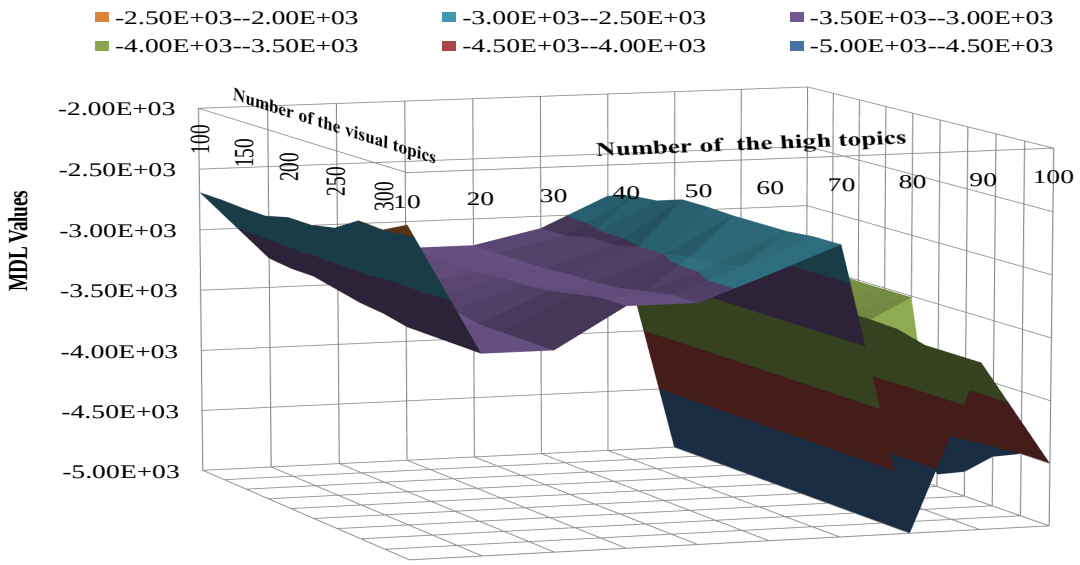

Figure 8.12: MDL values using the MIRFLICKR-25000 dataset.

8.2.3.3 Other parameters

Once the visual word vocabulary sizes and the number of latent topics of the different datasets are selected, we run empirical investigation to estimate

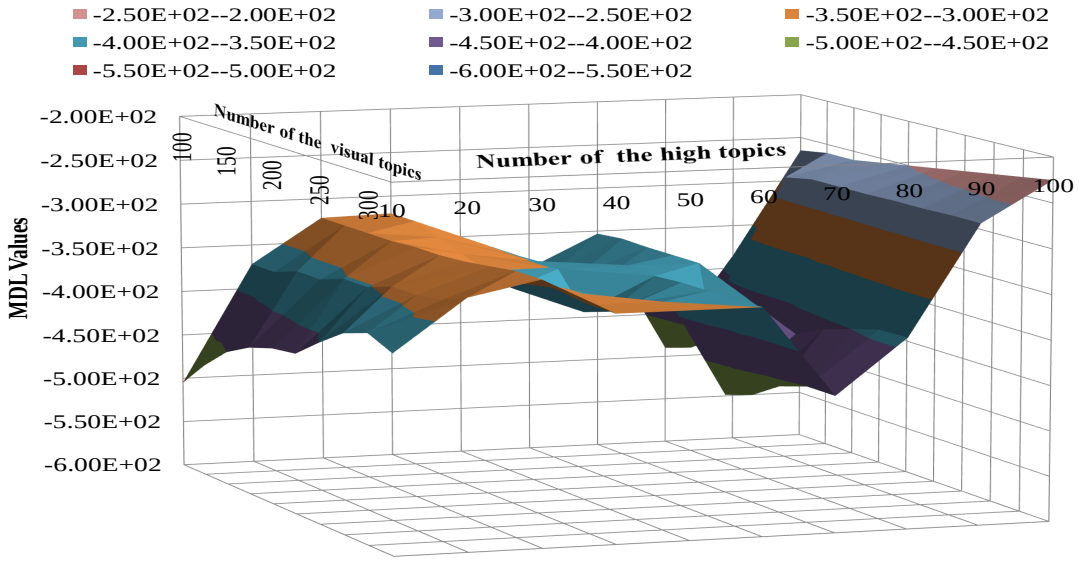

Figure 8.13: MDL values using the Caltech101.

Datasets	Number of the high latent topics	Number of the visual latent topics
NUS-WIDE	80	325
MIRFLICKR-25000	10	325
Caltech101	100	325

Table 8.3: Number of the high and visual latent topics as estimated by MDL for the three datasets.

further parameters. Table 8.4 shows the different values of the parameters for the different datasets.

Datasets	W	P	IVW	IVP	SS	SF	LS	WS
NUS-WIDE	3248	551	2750	500	$6.5\text{x}10^{-5}$	$1.5\text{x}10^{-3}$	$4.5\text{x}10^{-2}$	3.5×10^{-2}
MIRFLICKR-25000	1248	480	750	425	$4.5\text{x}10^{-3}$	$7.1\text{x}10^{-2}$	$9.5\text{x}10^{-2}$	0.1×10^{-2}
Caltech101	1480	409	1250	325	$0.1\text{x}\times10^{-3}$	0.2×10^{-2}	0.35	0.46

Table 8.4: Values of the different parameter settings.

Where:

- W is the SSVW vocabulary size.
- P is the SSVP vocabulary size.

– IVW is the SSIVW vocabulary size.

– IVP is the SSIVP vocabulary size.

– SS is the support threshold for the association rule mining theory.

– SF is the confidence threshold for the association rule mining theory.

– LS is the probability threshold for relevant visual latent topics.

– WS is the probability threshold for generating the SSVWS.

8.3 Assessment of the SSIVG representation performance in image retrieval

In this section, we study the performance of the proposed higher-level visual representation in retrieval using NUS-WIDE dataset.

We compare the performance of different representations: classical bag of visual words (BOW), the enhanced bag of visual words (E-BOW) that is introduced in Chapter 5, SSVW, SSVP, SSIVW, SSIVP and SSIVG that combine the SSIVW and the SSIVP representations. We also compare the performances of the visual glossaries generated from the pLSA and LDA models rather than the MSSA model, and we reference them here as SSIVG-pLSA and SSIVG-LDA representations, respectively.

We also extend the performance comparison to several other recently proposed higher-level representation methods specifically visual phrase pattern [171], descriptive visual glossary [174], and visual synset [177].

For all the representation methods, the traditional Vector Space Model of Information Retrieval is adapted using an inverted file structure and the $tf \times idf$

weighting for all representation except for the SSIVG representation, we use the proposed spatial weighting scheme and the $tf \times idf$ weighting as described in Section 7.5. In addition, the cosine distance is used for the similarity matching between the query image and the candidate images. The evaluation metric used for different experiments is the mean average precision (MAP) as described in Section 8.2.2.1.

8.3.1 Individual contributions of different representation levels in image retrieval

Figure 8.14 plots the mean average precisions for different representations in image retrieval. It is clear the E-BOW representation (MAP=0.193) outperforms the classical BOW representation (MAP=0.142). It is also obvious that SSIVW representation (MAP=0.225) is better than the E-BOW representation. The SSVW representation outperforms the BOW representation in the 81 categories except in 5 categories (*glacier*, *fire*, *sport*, *flags*, *sand*). We notice that the average number of classical visual words in these 5 categories is too small since the number of the detected interest points is too small. Having a small number of visual words leads to a fewer number of SSIVWs that are selected from the visual words, which affects the performance of the SSVW representation.

When considering only SSVPs (MAP=0.232), the performance is slightly better than that of SSVW (MAP=0.225). An SSVP representation contains both spatial and appearance information, which is assumed to be more informative than that of SSVW in many image categories. However, some query images in categories such as *sky* and *waterfall* do not present consistent spatial character-

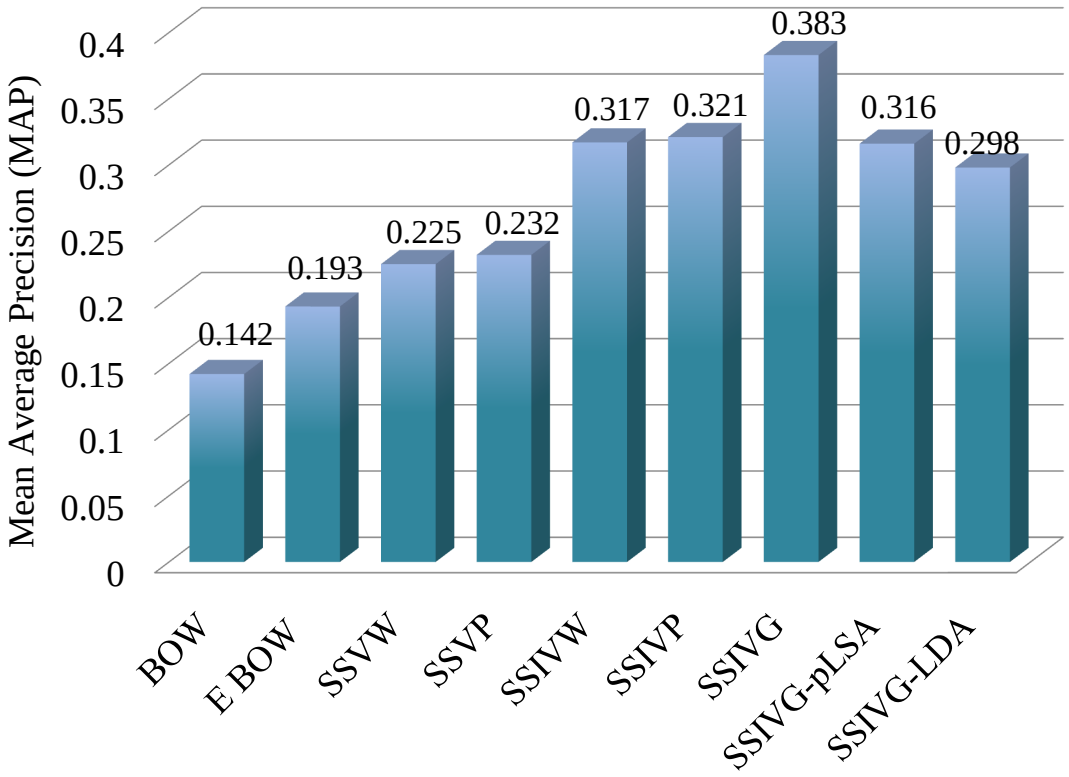

Figure 8.14: MAP results for the performance of BOW, E-BOW, SSVW, SSVP, SSIVW, SSIVP, SSIVG, SSIVG-pLSA, and SSIVG-LDA representations in image retrieval.

istics and contain very few or even zero SSVPs. Thus SSVPs do not work well for these cases.

The re-indexing of the SSVW and SSVP representations leads to the SSIVW and the SSIVP representation that have better performance (MAP=0.317 for the SSIVW representation and MAP=0.321 for the SSIVP representation). The combination of SSIVW and SSIVP into the SSIVG representation yields the best results with MAP=0.383. It also outperforms the SSIVG-pLSA (MAP=0.316) and SSIVG-LDA (MAP=0.298) representations especially in the categories that

have complicated visual scenes such as *weddings*, *military*, and *coral*.

8.3.2 Comparison of the SSIVG representation performance with other representation methods

Figure 8.15 shows the performance comparison between the SSIVG representation with visual phrase pattern, descriptive visual glossary, and visual synset. SSIVG representation performs better than others and the visual synset has the least performance (MAP=0.211) compared to others.

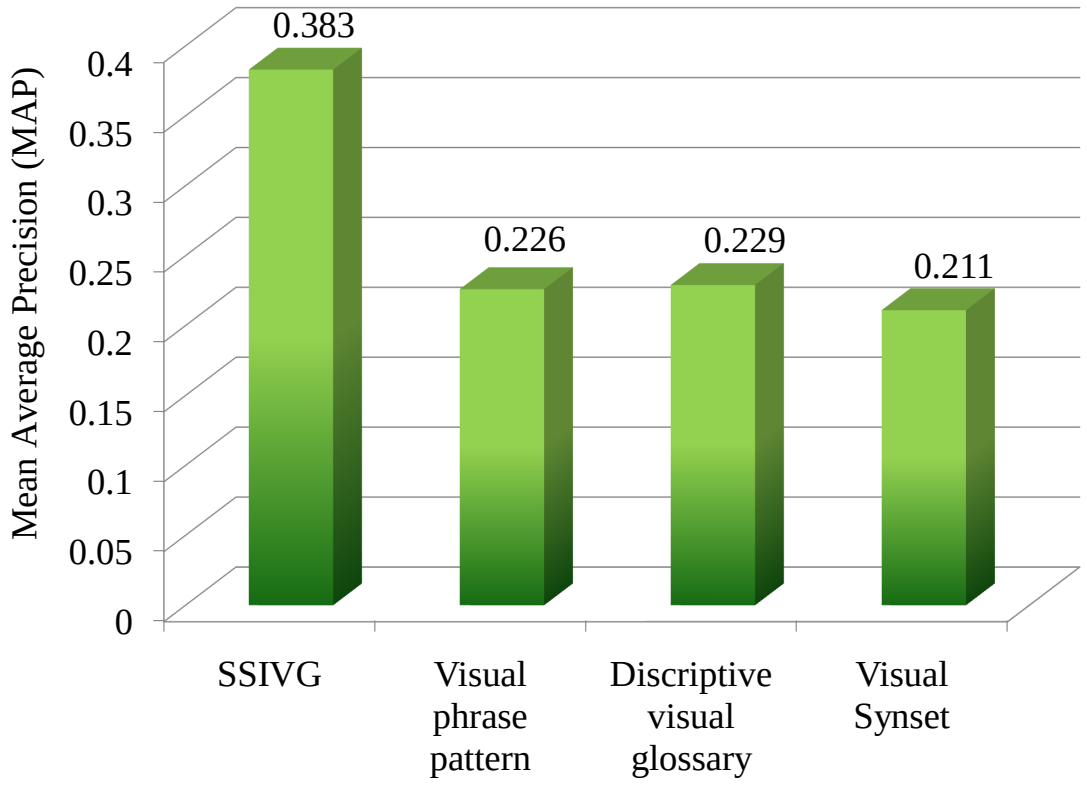

Figure 8.15: MAP results for different representations in image retrieval.

It is also noted that SSIVG representation outperforms the other representa-

tions in most of 81 classes while the visual phrase pattern representation outperforms SSIVG in only 3 categories (*dancing, train, computer*) and the descriptive visual glossary representation outperforms SSIVG in 2 categories (*fox, harbor*). Having this difference over a data set containing 81 categories and 269, 648 images emphasizes the good performance of the proposed representation.

8.4 Evaluation of the SSIVG Representation and MVBC Performance in Classification

In the following experiments, we study the performance of the SSIVG representation in classification using the vote-based classifier (MVBC). We test the proposed approach (SSIVG+MVBC) performance using MIRFLICKR-25000 data set. We also tested the proposed SSIVG representation using SVM with a linear kernel as a classifier. Again, we compare the classification performance of SSIVGS+MVBC with the other three higher-level visual representation (visual phrase pattern [171], descriptive visual glossary [177], and visual synset [177]) using SVM with a linear kernel as a classifier and $tf \times idf$ as weighting scheme.

Figure 8.16 plots the average classification precision results for each image class for different approaches.

It is clear that the proposed approach (SSIVG+MVBC) outperforms or performs closely to the SSIVG + SVM approach. SSIVG+MVBC approach also outperforms or performs equally comparing to other approaches. The highest classification performance is obtained in *sky*, and *sunset* classes. The different higher-level approaches perform well in these classes except the visual synset rep-

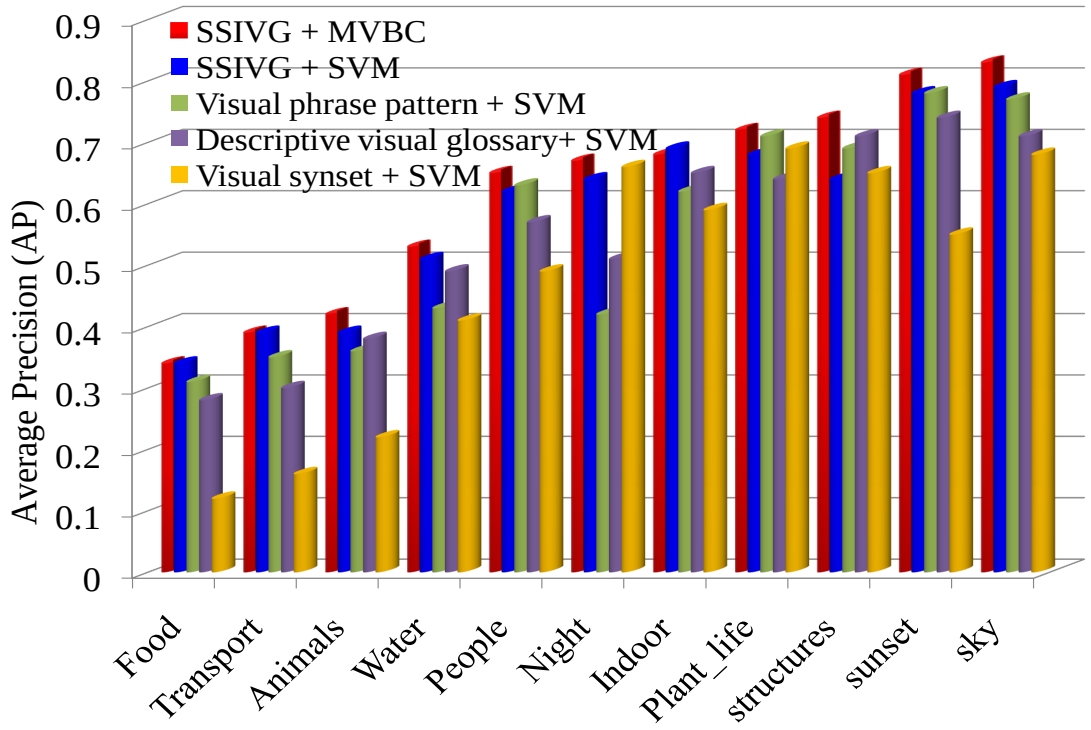

Figure 8.16: classification performance for different approaches.

resentation with SVM which perform worse than the other approaches. It is noted that all the images in both classes contain very specific colors and almost not so much texture. However, this is not always the case, for some *sky* images, there are cloudy skies or just a vague notion of sky somewhere in the images.

The least classification performances are in *animal, food*, and *transport* classes. Note that there is a wide variety of images that can be classified as containing *animal, food,* or *transport*. For example in the *animal* class, not only real animals that are clearly visible, but also hand drawn animals or parts of an animal result in the same class. In addition, in some images, the target object (*animal, food,* or *transport*) does not have to be the subject of the image, but it might also be

seen in the background. This makes the classification a challenging problem in these classes.

8.5 Assessment of the SSIVG representation performance in object recognition

Object recognition has been a popular research topic for many years. Many recently reported works show promising performance in this challenging recognition task. Since the SSIVGs effectively describe certain visual aspects (objects or scenes), it is straightforward that the SSIVGs in each object category should be discriminative for the corresponding object. Consequently, we utilize the object recognition task to illustrate the discriminative ability of SSIVGs.

We utilize the Caltech101 dataset for the object recognition task. For each test image, the training image category containing the same object is selected from the image database. In our approach, each test image is recognized by predicting the object class using the SSIVG representation and the MVBC. We compare this method with the visual phrase-based approach proposed by Zheng and Gao to retrieve images containing some desired objects. In this approach, each test image is recognized by computing the first 20 retrieved images in the training dataset.

Figure 8.17 shows the average precisions for the two approaches for each object category. We arrange the 101 classes from left to right with respect to the ascending order of average precisions of SSIVG representation in order to get a clearer representation.

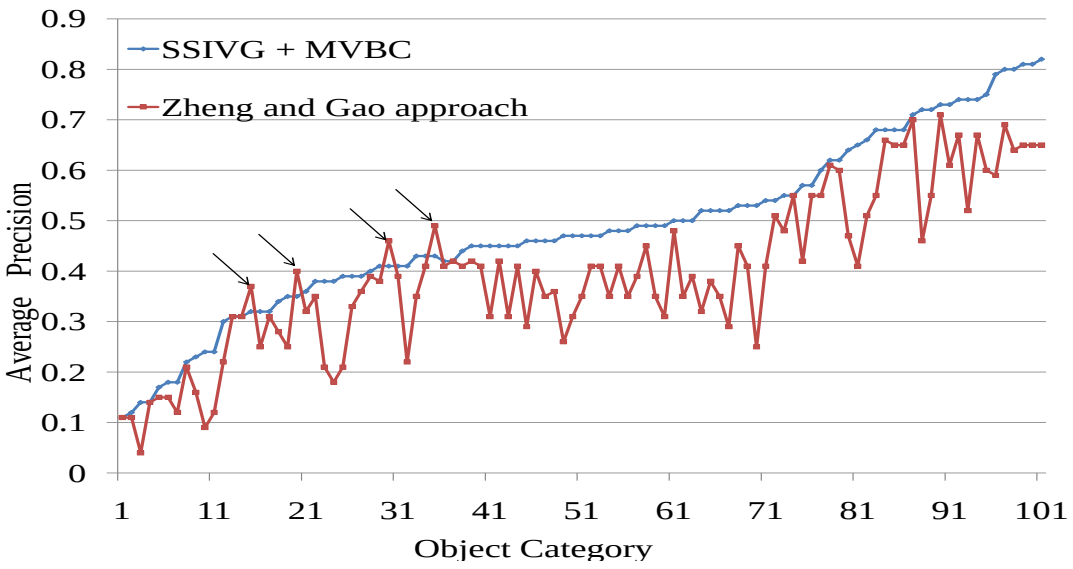

Figure 8.17: Object recognition performance for different approaches.

It is obvious from the results, that the proposed approach globally outperforms the other approach, except for four image classes (pyramid, revolver, dolphin, and stegosaurus) out of the 101 classes in the used data set.

8.6 Summary and conclusion

In this chapter we have presented the large-scale, extensive experimental studies that have demonstrated the good performance of the proposed SSIVG representation in image retrieval, image classification, and object recognition. We have compared the performance of the SSIVG representation to several recent approaches, specifically visual phrase pattern [171], descriptive visual glossary [174], and visual synset [177].

We have introduced the different criteria for estimating variety of parameter

settings, such as estimating the optimal visual word vocabulary sizes and the number of latent topics in the MSSA model. We have also examined the performance of the different layers of the proposed representation (Enhanced BOW (E-BOW), SSVW, SSVP, SSIVW, SSIVP, SSIVG). Consequently, we have extended the performance comparison to several other recently proposed higher-level representation methods in the image retrieval context that are mentioned above. Moreover, we have studied the performance of the proposed SSIVG representation and the proposed MVBC classifier in image classification. Finally, we have evaluated the performance of the proposed combination of the SSIVG representation and the MVBC classifier in the object recognition task.

Part IV

Conclusion and Future Work

Chapter 9

Conclusion

9.1 Summary

Due to the explosive spread of digital devices, the amount of digital content in terms of personal images grows rapidly. This increases the need for effective techniques for automatic processing, description, and structuring of large digital image archives. Most of the recent techniques are based on representing images using the text annotation associated with the images. Unreliable nature of the few tags assigned by users makes accurate tag based techniques infeasible.

This leads to increase the interest to image representation based on the visual content. Recently, it has been shown that the part-based representation especially bag of visual words (BOW) representation is more effective than the global representation. Indeed, one single image feature computed over the entire image is not sufficient to represent important local characteristics of different objects within the image.

Despite the good performance of BOW representation in different tasks such as image retrieval, scene classification, and object recognition, still there are drawbacks to be considered. This works aims to enhance the BOW representation and

propose a higher-level visual representation for semantic learning in large-scale image databases. The contributions of this thesis are three-fold as stated below.

1. An enhanced approach to construct the bag of visual words (BOW) representation is represented, which is the first level of representation in the proposed approach. The 5D color-spatial feature space are modeled for set of detected interest and edge points based on the Gaussian Mixture Model (GMM). Third, we extract at each interest point SURF local features. In addition to the SURF, we established a new local feature descriptor, Edge context, which plays a role as a descriptor complimentary to SURF descriptor. It describes, at each interest point, the distribution of the edge points that belongs to a given Gaussian cluster by returning to the 5D color-spatial space. Afterward, the two local feature vectors (SURF +Edge Context) are merged to get final local feature vectors. The quantization of the merged features into visual words is achieved by two clustering stages. A Hierarchical agglomerative clustering is performed to overcome the problem of the initial seed for the repeated k-means clustering that hierarchically partition the local feature space. This process results in the construction of a visual word vocabulary tree.

2. We propose a new probabilistic topic model: the Multilayer Semantically Significant (MSSA) model. The MSSA model studies the semantic inference of different atomic visual representation units (visual words) in order to select the semantically significant units. The MSSA model differs from other similar topic models by introducing two different latent topic layers, the high latent topics and visual latent topics that represent the high aspects

(images categories) and visual aspects (scenes, objects or part of the objects) of the images respectively. The KKT conditions are used to derive new multiplicative update rules in order to estimate the parameters of the MSSA model. In addition, the number of latent topics in the MSSA model is estimated using different mode selection criteria: the Akaike information criterion (AIC), the Bayesian information criterion (BIC), and the Minimum Description Length (MDL).

3. A higher-level visual representation is introduced: Semantically Significant Invariant Glossary (SSIVG) representation. This representation is based on the BOW representation and the MSSA model that are introduced above. We started by selecting the Semantically Significant Visual Words (SSVWs) from the visual words set in order to overcome the feature quantization noise. The selection process is based on the visual words semantic inferences that are estimated using the MSSA model. Subsequently, the discrimination power of the SSVWs are strengthened by building Semantically Significant Visual Phrases (SSVPs) from frequently co-occurring SSVW sets in the same local context, that are both involved in strong association rules, and semantically coherent. Moreover, the intra-class invariance power of the SSVWs and the SSVPs are boosted by a clustering based on their probability distributions to the relevant visual topics. These steps lead to form the Semantically Significant Invariant Visual Glossary (SSIVG) representation. Besides, we set up a new spatial weighting scheme dedicated to this representation is introduced with the vector space model for image retrieval and indexing. Moreover, we present a new Multilayer Vote-Based Classifier (MVBC) based on the SSIVG representation for image classification.

All the propositions made in this work have been implemented, and the successful results have been described and discussed. The implemented system is found to outperform various state-of-the-art image representation methods from the recent literature in three different contexts: image retrieval, image classification, and object recognition which validates the contributions.

9.2 Perspectives

Several directions are envisioned for future research based on the proposed higher-level visual representations as follows.

- *Parameters update*: As the large-scale online image repositories grow daily, an important aspect needs to be addressed when developing probabilistic topic models in future research. It will be essential to design on-line algorithms to continuously (re-)learn the parameters of the proposed MSSA model, as the content of digital databases is modified by the regular upload or deletion of images.
- *Invariance issue*: One of the major contributions of our work is to improve the invariance power of the BOW representation. The experimental results have shown that the proposed higher-level representation can partially bridge the visual differences between images of the same class and deliver a more coherent, invariant and compact representation of images. It will be interesting to investigate more on the invariance issue especially in the context large-scale databases where large intra-class variations can occur.
- *Abstract concepts*: The experimental evaluations used datasets consisting

mainly of objects and scene categories. With such datasets, the proposed approaches have shown to perform well. However, it is not clear to us how the performance would be affected when searching for abstract themes such as love, sad, celebrating, success, etc or landmark scenes. In future works this should be examined.

– *Video summarization*: Nowadays, users are facing an ever-increasing amount of television programs. The difficulty, however, is that the content of video programs is easily managed by the viewing devices. The existing video watching options for users are either to watch the whole video, fast forward to try and find the relevant portion, or to use electronic program guides (EPG) to get additional information. Video summarization is therefore essential to enable the user to view the content in different aspects. The proposed higher-level visual representation can be extended to video content. The extension can be based on cross-modal data (visual and textual closed captions contents), in order to reinforce the actual proposed representation that is based on one modality (only the visual content of images).

  Subsequently, a new generic framework of video summarization based on the extended higher-level semantic representation of video content can be designed. This framework will process the incoming video, extract and analyze closed caption text, determine the boundaries of program segments and commercial breaks, and extracts a program summary from a complete broadcast

Part V

Publications and Bibliography

Publications

[EMU+11] **Ismail El Sayad**, Jean Martinet, Thierry Urruty, Yassine Benabbas, and Chabane Dejraba. A semantically significant visual representation for social image retrieval. In *IEEE International Conference on Multimedia and Expo (ICME)*, page accepted, 2011.

[EMUD11] **Ismail El Sayad**, Jean Martinet, Thierry Urruty, and Chabane Dejraba. A semantic higher-level visual representation for object recognition. In *International conference on multimedia modeling (MMM)*, volume 6523 of *Lecture Notes in Computer Science*, pages 251-261, Springer-Verlag, 2011.

[ME11] Jean Martinet and **Ismail El Sayad**. Mid-level image descriptors. In *Intelligent Multimedia Databases and Information Retrieval: Advancing Applications and Technologies*, IGI Global, Hershey, 2011. ISBN13: 9781613501269, DOI: 10.4018/978-1-61350-126-9.

[EMU+10a] **Ismail El Sayad**, Jean Martinet, Thierry Urruty, Samir Amir, and Chabane Djeraba. Toward a higher-level visual representation for content-based image retrieval. In *ACM International Conference on Advances in Mobile Computing and Multimedia (ACM MoMM)*, pages 213-220, 2010.

[EMUD10a] **Ismail El Sayad**, Jean Martinet, Thierry Urruty, and Chabane Djeraba. A new spatial weighting scheme for bag-of-visual-words. In *IEEE International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 1-6, 2010.

[EMUD10b] **Ismail El Sayad**, Jean Martinet, Thierry Urruty, and Chabane Djeraba. Toward a higher-level visual representation for content-based image retrieval. *Multimedia Tools and Applications*, pages 1-28, 2010. 10.1007/s11042-010-0596-x.

[EMU+10b] **Ismail El Sayad**, Jean Martinet, Thierry Urruty, Samir Amir, and Chabane Djeraba. Effective object-based image retrieval using higher-level visual representation. In *IEEE International Conference on Machine and Web Intelligence (ICMWI)*, pages 218-224, 2010.

[EMU+10c] **Ismail El Sayad**, Jean Martinet, Thierry Urruty, Taner Danisman, Md. Haidar Sharif, and Chabane Djeraba. Using association rules and spatial weighting for an effective content based-image retrieval. In *International Conference on Computer Vision Theory and Applications (VISSAPP)*, volume 1, pages 112-117, 2010.

[ABD+10a] Samir Amir, Ioan Marius Bilasco, Taner Danisman, **Ismail El sayad**, and Chabane Djeraba. Multimedia metadata mapping: towards helping developers in their integration task. In *ACM International Conference on Advances in Mobile Computing and Multimedia (ACM MoMM)*, pages 205-212, 2010.

[ABD+10b] Samir Amir, Ioan Marius Bilasco, Taner Danisman, **Ismail El Sayad**, and Chabane Djeraba. Schema matching for integrating multimedia metadata. In *IEEE International Conference on Machine and Web Intelligence (ICMWI)*, pages 234-239, 2010.

[EMUD10c] **Ismail El Sayad**, Jean Martinet, Thierry Urruty, and Chabane Djeraba. Visual sentence-phrase-based document representation for effective and efficient content-based image retrieval. In *Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC)*, pages 157-162, 2010.

[UFL+10] Thierry Urruty, Yue Feng, Adel Lablack, Joemon M. Jose, and **Ismail El Sayad**. Classification et sélection de caractéristique basées sur les concepts sémantiques pour la recherche d'information multimédia. In *Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC)*, pages 85-90, 2010.

References

[1] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD Record*, volume 22, pages 207–216, 1993. 8, 121

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Int. Conf. Very Large Data Bases, VLDB*, volume 1215, pages 487–499. Citeseer, 1994. 122

[3] J.-H. Ahn, S.-K. Kim, J.-H. Oh, and S. Choi. A multiple nonnegative-matrix factorization of dynamic pet images. In *Asian Conference on Computer Vision*, 2004. 43

[4] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974. 106, 107

[5] J. Astola, P. Haavisto, and Y. Neuvo. Vector median filters. *Proceedings of the IEEE*, 78(4):678–689, 1990. 79

[6] R. Baeza-Yates, B. Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999. 57

[7] R. A. Baeza-Yates and B. A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999. 55, 57, 58

[8] L. D. Baker and A. McCallum. Distributional clustering of words for text classification. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 96–103. ACM, 1998. 132

[9] M. Barni, V. Cappellini, and A. Mecocci. Fast vector median filter based on Euclidean norm approximation. *IEEE Signal Processing Letters*, 1(6):92–94, 1994. 81

[10] H. Bay, A. Ess, T. Tuytelaars, and L. J. V. Gool. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3):346–359, 2008. vii, 72, 74, 75

[11] H. Bay, T. Tuytelaars, and L. J. V. Gool. Surf: Speeded up robust features. In *European Conference on Computer Vision (ECCV)*, volume 1, pages 404–417, 2006. 4, 20

[12] R. Bekkerman, R. El-Yaniv, N. Tishby, and Y. Winter. Distributional word clusters vs. words for text categorization. *J. Mach. Learn. Res.*, 3:1183–1208, 2003. 127, 129

[13] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(4):509–522, 2002. 87

[14] M. Berry, M. Browne, A. Langville, V. Pauca, and R. Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis*, 52(1):155–173, 2007. 43

[15] J. Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. Technical Report TR-97-021, ICSI, 1997. 72, 82

[16] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003. 7, 32, 35, 36, 37

[17] P. Brodatz. *Textures: A Photographic Album for Artists and Designers*. Dover Publications, August 1999. 17

[18] M. Brown and D. Lowe. Invariant features from interest point groups. In *British Machine Vision Conference, Cardiff, Wales*, pages 656–665, 2002. 75

[19] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *the National Academy of Sciences of the United States of America*, 101(12):4164–4169, March 2004. 43

[20] W. L. Buntine. Operations for learning with graphical models. *J. Artif. Intell. Res. (JAIR)*, 2:159–225, 1994. 39

[21] W. L. Buntine. Variational extensions to em and multinomial pca. In *European Conference on Machine Learning (ECML)*, volume 2, pages 23–34. Springer-Verlag, 2002. 42

[22] W. L. Buntine. Variational extensions to em and multinomial pca. In *European Conference on Machine Learning (ECML)*, pages 23–34, 2002. 46

[23] C. Burges and B. Schšolkopf. Improving the accuracy and speed of support vector machines. *Advances in neural information processing systems*, pages 375–381, 1997. 5, 86

[24] J. Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(6):679–698, November 1986. 72, 76

[25] J. Carlson, M. Muguira, J. Jordan, G. Flachs, and A. Peterson. Final report: Weighted neighbor data mining. Technical report, Sandia National Labs., Albuquerque, NM (US), 2000. 54

[26] C. Carson, M. Thomas, S. Belongie, J. M. Hellerstein, and J. Malik. Blobworld: A system for region-based image indexing and retrieval. In *VISUAL*, pages 509–516, 1999. 14, 26, 30, 55, 56

[27] S. K. Chang, Q. Y. Shi, and C. W. Yan. Iconic indexing by 2-d strings. *IEEE Trans. Pattern Anal. Mach. Intell.*, 9:413–428, May 1987. 5, 22

[28] Y.-C. Cho and S. Choi. Nonnegative features of spectro-temporal sounds for classification. *Pattern Recogn. Lett.*, 26:1327–1336, July 2005. 43

[29] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng. Nus-wide: A real-world web image database from national university of singapore. In *ACM International Conference on Image and Video Retrieval (CIVR)*. 145

[30] A. Cichocki, S.-i. Amari, R. Zdunek, R. Kompass, G. Hori, and Z. He. Extended smart algorithms for non-negative matrix factorization. In *Artificial Intelligence and Soft Computing*, volume 4029 of *Lecture Notes in Computer Science*, pages 548–562. Springer Berlin Heidelberg, 2006. 44

[31] D. A. Cohn and T. Hofmann. The missing link - a probabilistic model of document content and hypertext connectivity. In *NIPS*, pages 430–436, 2000. 38

[32] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. New York: Wiley, 1991. 130

[33] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 886–893, 2005. 120

[34] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990. 34

[35] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal statistical society, series B*, 39(1):1–38, 1977. 82, 100

[36] I. S. Dhillon, S. Mallela, and R. Kumar. A divisive information-theoretic feature clustering algorithm for text classification. *Journal of Machine Learning Research*, 3:1265–1287, 2003. 130

[37] I. S. Dhillon and S. Sra. Generalized nonnegative matrix approximations with bregman divergences. In *NIPS*, 2005. 44

[38] I. El Sayad, J. Martinet, T. Urruty, S. Amir, and C. Djeraba. Toward a higher-level visual representation for content-based image retrieval. In *ACM International Conference on Advances in Mobile Computing and Multimedia (ACM MoMM)*, pages 213–220, 2010. 118

[39] I. El Sayad, J. Martinet, T. Urruty, and C. Dejraba. A semantic higher-level visual representation for object recognition. In *International conference on multimedia modeling (MMM), volume 6523 of Lecture Notes in Computer Science*, pages 251–261. Springer Berlin / Heidelberg, 2011. 96

[40] I. El Sayad, J. Martinet, T. Urruty, and C. Djeraba. A new spatial weighting scheme for bag-of-visual-words. In *IEEE International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 1–6, 2010. 87

[41] I. El Sayad, J. Martinet, T. Urruty, and C. Djeraba. Toward a higher-level visual representation for content-based image retrieval. *Multimedia Tools and Applications*, pages 1–28, 2010. 82, 87, 90

[42] I. Elsayad, J. Martinet, T. Urruty, T. Danisman, M. H. Sharif, and C. Djeraba. Using association rules and spatial weighting for an effective content based-image retrieval. In *VISAPP (1)*, pages 112–117, 2010. 82

[43] F. Ennesser and G. G. Medioni. Finding waldo, or focus of attention using local color information. *IEEE Trans. Pattern Anal. Mach. Intell.*, 17(8):805–809, 1995. 24

[44] C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic, and W. Equitz. Efficient and effective querying by image content. *J. Intell. Inf. Syst.*, 3(3-4):231–262, 1994. 14, 30

[45] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70, 2007. 145

[46] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 524–531. IEEE Computer Society, 2005. 26

[47] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. Query by image and video content: The qbic system. *Computer*, 28(9):23–32, 1995. 25

[48] G. Furnas, S. Deerwester, S. Dumais, T. Landauer, R. Harshman, L. Streeter, and K. Lochbaum. Information retrieval using a singular value decomposition model of latent semantic structure. In *the 11th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 465–480. ACM, 1988. 57

[49] S. Gao, I. Tsang, L.-T. Chia, and P. Zhao. Local features are not lonely - sparse coding for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3555 –3561, jun. 2010. 28

[50] E. Gaussier and C. Goutte. Relation between plsa and nmf and implications. In *the annual international ACM SIGIR conference on Research and development in information retrieval(SIGIR)*, pages 601–602, 2005. 46, 100

[51] T. Gevers and A. W. M. Smeulders. Content-based image retrieval by viewpoint-invariant color indexing. *Image Vision Comput.*, 17(7):475–488, 1999. 15

[52] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1458–1465, 2005.

[53] T. L. Griffiths and M. Steyvers. A probabilistic approach to semantic representation, 2002. 35, 37, 40

[54] T. L. Griffiths and M. Steyvers. Finding scientific topics. *the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, April 2004.

[55] A. Gupta and R. Jain. Visual information retrieval. *Commun. ACM*, 40(5):70–79, 1997. 24

[56] J. Han and M. Kamber. *Data mining: concepts and techniques*. Morgan Kaufmann, 2006. 122

[57] J. Hawkins and S. Blakeslee. *On intelligence*. Owl Books, 2005. 98

[58] E. Hoerster, R. Lienhart, and M. Slaney. Image retrieval on large-scale image databases. In *ACM international conference on Image and video retrieval (CIVR)*, pages 17–24, 2007. 2

[59] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1/2):177–196, 2001. vii, 7, 32, 35, 41

[60] N. V. Hoíng, V. Gouet-Brunet, M. Rukoz, and M. Manouvrier. Embedding spatial information into image content description for scene retrieval. *Pattern Recogn.*, 43(9):3013–3024, 2010. 29

[61] P. O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, 2004. 44

[62] K. A. Hua, K. Vu, and J.-H. Oh. Sammatch: a flexible and efficient sampling-based image retrieval technique for large image databases. In *ACM Multimedia (1)*, pages 225–234, 1999. 15, 16

[63] J. Huang, R. Kumar, M. Mitra, and W.-J. Zhu. Spatial color indexing and applications. In *IEEE International Conference on Computer Vision (ICCV)*, pages 602–607, 1998. 25

[64] M. J. Huiskes and M. S. Lew. The mir flickr retrieval evaluation. In *ACM International Conference on Multimedia Information Retrieval (ACM MIR)*. ACM, 2008. 145

[65] N. Ikonomakis, K. N. Plataniotis, and A. N. Venetsanopoulos. Color image segmentation for multimedia applications. *Journal of Intelligent and Robotic Systems*, 28(1-2):5–20, 2000. 15

[66] P. Jaccard. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. 1901. 65

[67] A. K. Jain and A. Vailaya. Image retrieval using color and shape. *Pattern Recognition*, 29(8):1233–1244, 1996. 24

[68] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, 2010. 26

[69] J. Jeon and R. Manmatha. Using maximum entropy for automatic image annotation. In *ACM International Conference on Image and Video Retrieval (CIVR)*, pages 24–32, 2004. 26, 55, 56, 96

[70] Y. G. Jiang, C. W. Ngo, and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *ACM International Conference on Image and Video Retrieval (CIVR)*, pages 494–501. ACM, 2009. 26, 66, 134

[71] F. Jing, M. Li, L. Z. 0001, H. Zhang, and B. Zhang. Learning in region-based image retrieval. In *ACM International Conference on Image and Video Retrieval (CIVR)*, pages 206–215, 2003. 15

[72] F. Jing, M. Li, L. Zhang, H. jiang Zhang, and B. Zhang. Learning in region-based image retrieval. In *in Proceedings of the IEEE International Symposium on Circuits and Systems*, pages 206–215. Springer, 2003. 14, 27, 30, 55, 56, 58

[73] I. T. Jolliffe. *Principal Component Analysis*. Springer, second edition, 2002. 57

[74] K. S. Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21, 1972. 59

[75] R. Jörnsten and B. Yu. Simultaneous gene clustering and subset selection for sample classification via MDL. *Bioinformatics*, 19(9):1100, 2003. 110

[76] M. Kim and S. Choi. Monaural music source separation: Nonnegativity, sparseness, and shift-invariance. In *Independent Component Analysis and Blind Signal Separation*, volume 3889 of *Lecture Notes in Computer Science*, pages 617–624. Springer Berlin Heidelberg, 2006. 43

[77] J. Koenderink. The structure of images. *Biological cybernetics*, 50(5):363–370, 1984. 74

[78] R. Kompass. A generalized divergence measure for nonnegative matrix factorization. *Neural Comput.*, 19:780–791, March 2007. 44

[79] R. Krishnapuram, A. Joshi, and L. Yi. A fuzzy relative of the k-medoids algorithm with application to web document and snippet clustering. In *IEEE International Conference Fuzzy Systems (FUZZI)*, volume 3, pages 1281–1286, 2002. 57

[80] H. W. Kuhn. Nonlinear programming: a historical view. *SIGMAP Bull.*, pages 6–18, 1982. 96, 101

[81] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 2169–2178, 2006. 26, 27, 150

[82] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 5, 86

[83] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, October 1999. 42, 44

[84] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *NIPS*, pages 556–562, 2000. 42, 44

[85] J. S. Lee, D. D. Lee, S. Choi, and D. S. Lee. Application of non-negative matrix factorization to dynamic positron emission tomography. In *In 3rd International Conference on Independent Component Analysis and Blind Signal Separation*, pages 629–632, 2001. 42, 43

[86] R. Lienhart, S. Romberg, and E. Hörster. Multilayer plsa for multimodal image retrieval. In *ACM International Conference on Image and Video Retrieval (CIVR)*, page 9. ACM, 2009. 39, 98

[87] J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–, 1991. 130

[88] F. Liu and R. W. Picard. Periodicity, directionality, and randomness: Wold features for image modeling and retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(7):722–733, 1996. 17

[89] Y. Liu, D. Zhang, G. Lu, and W. Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40(1):262–282, 2007. 4, 14, 85

[90] Y. Liu, D. Zhang, G. Lu, and W. Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40(1):262–282, 2007. 3

[91] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma. Region-based image retrieval with perceptual colors. In *Pacific-Rim Conference on Multimedia (PCM)*, volume 2, pages 931–938, 2004. 15

[92] D. Lowe. Object recognition from local scale-invariant features. In *IEEE International Conference on Computer Vision (ICCV)*, page 1150. Published by the IEEE Computer Society, 1999. 4

[93] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 19, 74, 84

[94] H. Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development*, 1(4):309–317, 2010. 58

[95] R. Lukac, B. Smolka, K. Martin, K. Plataniotis, and A. Venetsanopoulos. Vector filtering for color imaging. *IEEE Signal Processing Magazine*, 2005. 16

[96] W. Ma and B. Manjunath. Netra: a toolbox for navigating large image databases. In *ICIP*, 1997. 16, 17, 21

[97] W. R. Maneesha, W. Ren, M. Singh, and S. Singh. Image retrieval using spatial context. In *Ninth International Workshop on Systems, Signals and Image Processing (IWSSIP)*, 2002. 22

[98] B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada. Color and texture descriptors. *IEEE Trans. Circuits Syst. Video Techn.*, 11(6):703–715, 2001. 15, 17

[99] C. Manning and H. Schütze. Foundations of statistical natural language processing. *Computational Linguistics*, 19(2). 148

[100] J. Martinet, Y. Chiaramella, and P. Mulhem. A relational vector space model using an advanced weighting scheme for image retrieval. *Inf. Process. Manage.*, 47(3):391–414, 2011.

[101] J. Martinet and S. Satoh. A study of intra-modal association rules for visual modality representation. In *International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 344 –350, 2007. 123

[102] R. Mehrotra and J. E. Gary. Similar-shape retrieval in shape data management. *IEEE Computer*, 28(9):57–62, 1995. 20

[103] V. Mezaris, I. Kompatsiaris, and M. G. Strintzis. An ontology approach to object-based image retrieval. In *IEEE International Conference on Image Processing (ICIP)*, volume 2, pages 511–514, 2003. 15, 16, 21

[104] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *IEEE International Conference on Computer Vision (ICCV)*, pages 525–531, 2001. 74

[105] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004. 26

[106] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(10):1615–1630, 2005. 19, 26

[107] B. Moghaddam, T. Jebara, and A. Pentland. Bayesian face recognition. *Pattern Recognition*, 33(11):1771–1782, 2000. 5, 86

[108] A. Mojsilovic, J. Gomes, and Rogowitz. ISee: perceptual features for image library navigation. 2001. 21

[109] A. Mojsilovic and B. Rogowitz. Capturing image semantics with low-level descriptors. volume 1, pages 18 –21 vol.1, 2001. 22

[110] F. Mokhtarian and S. Abbasi. Shape similarity retrieval under affine transforms. *Pattern Recognition*, 35(1):31–41, 2002. 21

[111] C. Morand, J. Benois-Pineau, J. P. Domenger, J. Zepeda, E. Kijak, and C. Guillemot. Scalable object-based video retrieval in hd video databases. *Image Commun.*, 25(6):450–465, 2010. 26

[112] H. Murase and S. Nayar. Visual learning and recognition of 3-D objects from appearance. *International journal of computer vision*, 14(1):5–24, 1995. 86

[113] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 2161–2168, 2006. 90

[114] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(7):971–987, 2002. 19

[115] P. Paatero and U. Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994. 42

[116] K. Papineni. Why inverse document frequency? In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, NAACL '01, pages 1–8. Association for Computational Linguistics, 2001. 59

[117] G. Pass and R. Zabih. Comparing images using joint histograms. *Multimedia Syst.*, 7:234–240, May 1999. 24

[118] K. Plataniotis, D. Androutsos, and A. Venetsanopoulos. Adaptive fuzzy systems for multichannel signal processing. *Proceedings of the IEEE*, 87(9):1601 –1622, 1999. 16

[119] K. N. Plataniotis and A. N. Venetsanopoulos. *Color image processing and applications*. Springer-Verlag New York, Inc., 2000. 16

[120] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, and T. Tuytelaars. A thousand words in a scene. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(9):1575–1589, 2007. 26, 55, 56, 58, 96

[121] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, and T. Tuytelaars. A thousand words in a scene. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(9):1575–1589, Sept. 2007.

[122] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. J. V. Gool. Modeling scenes with local descriptors and latent aspects. pages 883–890, 2005. 26, 55, 56, 96

[123] A. Rao, R. K. Srihari, and Z. Zhang. Geometric histogram: a distribution of geometric configurations of color subsets. volume 3964, pages 91–101. SPIE, 1999. 25

[124] J. Rissanen. *Stochastic Complexity in Statistical Inquiry Theory*. World Scientific Publishing Co., Inc., 1989. 109

[125] S. Robertson. Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation*, 60(5):503–520, 2004. 59, 60, 61, 66

[126] P. Salembier and T. Sikora. *Introduction to MPEG-7: Multimedia Content Description Interface*. John Wiley & Sons, Inc., 2002. 15, 17, 18, 21

[127] G. Salton. *The SMART Retrieval System—Experiments in Automatic Document Processing*. Prentice-Hall, Inc., 1971. 65

[128] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988. 54, 57

[129] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, 1975. 55, 136

[130] L. K. Saul and D. D. Lee. Multiplicative updates for classification by mixture models. In *NIPS*, pages 897–904, 2001. 42

[131] G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978. 106, 108

[132] I. K. Sethi, I. L. Coman, and D. Stan. Mining association rules between low-level image features and high-level concepts. volume 4384, pages 279–290. SPIE, 2001. 13

[133] F. Shahnaz, M. W. Berry, V. P. Pauca, and R. J. Plemmons. Document clustering using nonnegative matrix factorization. *Inf. Process. Manage.*, 42:373–386, March 2006. 43, 44

[134] C. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001. 54

[135] M. Shashanka, B. Raj, and P. Smaragdis. Probabilistic Latent Variable Models as nonnegative factorizations. *Computational Intelligence and Neuroscience*, pages 1–8, 2008. vii, 46, 48, 50

[136] R. Shi, H. Feng, T.-S. Chua, and C.-H. Lee. An adaptive image content representation and segmentation approach to automatic image annotation. In *ACM International Conference on Image and Video Retrieva (CIVR)*, pages 545–554, 2004. 15

[137] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1470–1477, 2003. 3, 26, 27, 30, 150

[138] J. Sivic and A. Zisserman. Video data mining using configurations of viewpoint invariant regions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 488–495, 2004. 120

[139] N. Slonim and N. Tishby. The power of word clusters for text classification. In *In 23rd European Colloquium on Information Retrieval Research*, 2001. 132

[140] A. W. M. Smeulders, S. Member, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:1349–1380, 2000. 13, 23

[141] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, 2000. 2

[142] J. Smith and Chung-Sheng-Li. Decoding image semantics using composite region templates. In *IEEE Workshop on Content-Based Access of Image and Video Libraries, 1998.*, 1998. 22

[143] J. R. Smith and S.-F. Chang. Visualseek: a fully automated content-based image query system. In *ACM Multimedia (ACM MM)*, MULTIMEDIA '96, pages 87–98, 1996. 14, 30

[144] Y. Song, W. Wang, and A. Zhang. Automatic annotation and retrieval of images. *World Wide Web*, 6:209–231, 2003. 10.1023/A:1023674722438. 21

[145] D. M. Squire, W. Müller, H. Müller, and J. Raki. Content-based query of image databases, inspirations from text retrieval: Inverted files, frequency-based weights and relevance feedback. In *Pattern Recognition Letters*, pages 143–149, 1999. 14, 30

[146] P. Stanchev. Using image mining for image retrieval. In *IASTED International Conference on Computer Science and Technology*, pages 214–218, 2003. 16

[147] P. L. Stanchev, D. Green, and B. Dimitrov. High level color similarity retrieval. *International Journal Information Theories and Applications*, 10:283–287, 2003. 16

[148] M. Steyvers and T. Griffiths. *Probabilistic Topic Models*. Lawrence Erlbaum Associates, 2007. vii, 33, 37

[149] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths. Probabilistic author-topic models for information discovery. In *ACM SIGKDD International Conference on Knowledge discovery and data mining KDD*, pages 306–315. ACM, 2004. 38

[150] M. Stokes, M. Anderson, S. Chandrasekar, and R. Motta. A standard default color space for the internet-sRGB. *Microsoft and Hewlett-Packard Joint Report*, 1996. viii, 72, 79, 80

[151] M. Stricker and M. Swain. The capacity of color histogram indexing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 704–708, 1994. 24

[152] M. J. Swain and D. H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991. 24

[153] L. S.Z., X. W. Hou, H. J. Zhang, and Q. S. Cheng. Learning spatially localized, parts-based representation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 207–212, 2001. 43

[154] V. Tahani. A fuzzy model of document retrieval systems. *Information Processing & Management*, 12(3):177–187, 1976. 55

[155] H. Tamura, S. Mori, and T. Yamawaki. Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man and Cybernetics*, 8(6):460–473, 1978. 17

[156] C. P. Town and D. Sinclair. Content-based image retrieval using semantic visual categories, 2001. 15, 21

[157] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, 2 edition, 1979. 56, 58, 136

[158] M. Varma and A. Zisserman. Texture classification: Are filter banks necessary? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 691–698, 2003. 19

[159] P. A. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 84

[160] J. Z. Wang, J. Li, D. Chan, and G. Wiederhold. Semantics-sensitive retrieval for digital picture libraries. *Digital Library Magazine*, 5(11), 1999. 15, 21

[161] J. Z. Wang, J. Li, and G. Wiederhold. Simplicity: Semantics-sensitive integrated matching for picture libraries. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(9):947–963, 2001. 16, 17

[162] W. Wang, Y. Song, and A. Zhang. Semantics-based image retrieval by region saliency. In *ACM International Conference on Image and Video Retrieval (CIVR)*, pages 29–37, 2002. 15

[163] R. Weiss, J. Bello, R. Weiss, J. B. (marl, R. Weiss, J. B. (marl, and F. B. synchronous Chroma Features [ellis. Shift-invariant probabilistic latent component analysis. *Journal ofMachine Learning Research*, 2010. 47, 49

[164] I. H. Witten, A. Moffat, and T. C. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images, Second Edition.* Morgan Kaufmann, 1999. 137

[165] S. L. W.K. Leow. Scale and orientation-invariant texture matching for image retrieval. In *M.K. Pietikainen (Ed.), Texture Analysis in Machine Vision, World Scientific*, pages 1–13. 17

[166] Z. Wu, Q. Ke, M. Isard, and J. Sun. Bundling features for large scale partial-duplicate web image search. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25–32, 2009. 28

[167] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *ACM SIGIR*, pages 267–273. ACM, 2003. 43, 44

[168] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo. Evaluating bag-of-visual-words representations in scene classification. In *ACM Multimedia Information Retrieval(ACM MIR)*, pages 197–206, 2007. 3, 27

[169] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1794 –1801, 2009. 27

[170] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *International Conference on Machine Learning (ICML)*, pages 412–420, 1997. 66, 134

[171] J. Yuan, Y. Wu, and M. Yang. Discovery of collocation patterns: from visual words to visual phrases. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2007. 28, 159, 163, 166

[172] J. Yuan, Y. Wu, and M. Yang. Discovery of collocation patterns: from visual words to visual phrases. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007. 28, 55, 56, 96, 119

[173] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision*, 73(2):213–238, 2007. 27, 55, 56, 58

[174] S. Zhang, Q. Tian, G. Hua, Q. Huang, and S. Li. Descriptive visual words and visual phrases for image applications. In *ACM Multimedia(ACM MM)*, pages 75–84, 2009. 28, 30, 55, 56, 58, 96, 159, 166

[175] Q.-F. Zheng and W. Gao. Constructing visual phrases for effective and efficient object-based image retrieval. *TOMCCAP*, 5(1), 2008. 28, 55, 56, 58, 96

[176] X. Zheng, D. Cai, X. He, W.-Y. Ma, and X. Lin. Locality preserving clustering for image database. In *ACM Multimedia*, pages 885–891, 2004. 15

[177] Y. T. Zheng, M. Zhao, S.-Y. Neo, T.-S. Chua, and Q. Tian. Visual synset: Towards a higher-level visual representation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008. 29, 55, 56, 96, 119, 159, 163, 166

[178] X. Zhou, X. Zhuang, S. Yan, S.-F. Chang, M. Hasegawa-Johnson, and T. S. Huang. Sift-bag kernel for video event analysis. In *ACM Multimedia (ACM MM)*, pages 229–238, 2008. 26

[179] X. S. Zhou and T. S. Huang. Cbir: From low-level features to highlevel semantics. In *International Conference on Image and Video Communication and Processing (SPIE)*, pages 24–28, 2000. 22, 23