Contribution à la détection de concepts sur des images utilisant des descripteurs visuels et textuels

Yu Zhang

THESE

pour obtenir le grade de
DOCTEUR DE L'ECOLE CENTRALE DE LYON
Spécialisté: Informatique

dans le cadre de l'Ecole Doctorale InfoMaths
présentée et soutenue

Contribution to Concept Detection on Images Using Visual and Textual Descriptors

---

Zhang Yu
Mars 2014

Directeur de thèse: Liming CHEN
Co-directeur de thèse: Stéphane BRES

JURY

Pr. RAMEL Jean Yves	LI - Tours	Rapporteur
DR. QUENOT Georges	LIG - Grenoble	Rapporteur
Pr.VINCENT Nicole	LIPADE - Paris	Présidente
MCF. MAHDI Walid	MIRACL - Sfax	Examinateur
Pr. Liming CHEN	Ecole Centrale de Lyon	Directeur de thèse
MCF. Stephane BRES	INSA Lyon	Co-directeur de thèse

Contents

Abstract

This thesis is dedicated to the problem of training and integration strategies of several modalities (visual, textual), in order to perform an efficient Visual Concept Detection and Annotation (VCDA) task, which has become a very popular and important research topic in recent years because of its wide range of application such as image/video indexing and retrieval, security access control, video monitoring, etc. Despite a lot of efforts and progress that have been made during the past years, it remains an open problem and is still considered as one of the most challenging problems in computer vision community, mainly due to inter-class similarities and intra-class variations like occlusion, background clutter, changes in viewpoint, pose, scale and illumination. This means that the image content can hardly be described by low-level visual features. In order to address these problems, the text associated with images is used to capture valuable semantic meanings about image content. Moreover, In order to benefit from both visual models and textual models, we propose multimodal approach. As the typical visual models, designing good visual descriptors and modeling these descriptors play an important role. Meanwhile how to organize the text associated with images is also very important. In this context, the objective of this thesis is to propose some innovative contributions for the task of VCDA. For visual models, a novel visual features/descriptors was proposed, which effectively and efficiently represent the visual content of images/videos. In addition, a novel method for encoding local binary descriptors was present. For textual models, we proposed two kinds of novel textual descriptor. The first descriptor is semantic Bag-of-Words(sBoW) using a dictionary. The second descriptor is Image Distance Feature(IDF) based on tags associated with images. Finally, in order to benefit from both visual models and textual models, fusion is carried out by MKL efficiently embed.

Firstly, we present a novel method for encoding local binary descriptors for

Visual Object Categorization (VOC). Nowadays, local binary descriptors, e.g. LBP and BRIEF, have become very popular in image matching tasks because of their fast computation and matching using binary bitstrings. However, the bottleneck of applying them in the domain of VOC lies in the high dimensional histograms produced by encoding these binary bitstrings into decimal codes. To solve this problem, we propose to encode local binary bitstrings directly by the Bag-of-Features (BoF) model with Hamming distance. The advantages of this approach are two-fold: (1) It solves the high dimensionality issue of the traditional binary bitstring encoding methods, making local binary descriptors more feasible for the task of VOC, especially when more bits are considered; (2) It is computationally efficient because the Hamming distance, which is very suitable for comparing bitstrings, is based on bitwise XOR operations that can be fast computed on modern CPUs. The proposed method is validated by applying on LBP feature for the purpose of VOC.

Secondly, we propose a novel representation, called sampled multi-scale color Local Binary Pattern (SMC-LBP), and apply it to Visual Object Classes (VOC) Recognition. The Local Binary Pattern (LBP) has been proven to be effective for image representation, but it is too local to be robust. Meanwhile such a design cannot fully exploit the discriminative capacity of the features available and deal with various changes in lighting and viewing conditions in real-world scenes. In order to address these problems, we propose SMC-LBP, which randomly samples the neighboring pixels across different scale circles, instead of pixels from individual circular in the original LBP scheme. The proposed descriptor presents several advantages: (1) It encodes not only single scale but also multiple scales of image patterns, and hence provides a more complete image information than the original LBP descriptor; (2) It cooperates with color information, therefore its photometric invariance property and discriminative power is enhanced.

Thirdly, we present two kinds of methods for building textual feature defined on semantic distance based on *Wordnet* distance for Visual Concept Detection and Annotation(VCDA). Nowadays, the tags associated with images have been popularly used in the VCDA task, because they contain valuable information about image content that can hardly be described by low-level visual features. Traditionally the

term frequencies model is used to capture this useful text information. However, the shortcoming in the term frequencies model lies that the valuable semantic information cannot be captured. To solve this problem, we propose two kinds of features. Firstly, we proposed two methods to associate to images a signature computed from the textual information. The first one uses a dictionary and is able to treat situation where textual information is huge (text associated to images on web pages for instance). For this method, the advantages of this approach are two-fold: (1) It can capture tags semantic information that is hardly described by the term frequencies model. (2) It solves the high dimensionality issue of the codebook vocabulary construction, reducing the size of the tags representation. The second one does not need any dictionary but is only usable in situations where textual information is reduced to a set of few tags. For this method, besides the advantages of the first method, the second method is more robust because it dose not rely on a dictionary construction.

Finally, we present a multimodal framework for Visual Concept Detection and Annotation(VCDA) task based on Multiple Kernel Learning(MKL), To extract discriminative visual features and build visual kernels. Meanwhile the tags associated with images are used to build the textual kernels. Finally, in order to benefit from both visual models and textual models, fusion is carried out by MKL efficiently embed. This integration strategy based on the multimodel framework obtains superior performance compared with single-model such as visual models or textual models.

Chapter 1

Introduction

Contents

1.1 Research Topic

With the rapid popularization of the digital cameras and smart phones, more and more images are shared in the internet. This means more and more information around us is compound of text-based and multimedia-based, especially in the form of images and videos associated with text. For example, the very famous online photo sharing website Flickr reported in August 2011 that it was hosting more than 6 billion photos already and that more than 3.5 million new images were uploaded daily.[1] Another famous social networking website Facebook announced in October 2011 that it was hosting about 140 billion images and thus becomes the largest

[1] http://en.wikipedia.org/wiki/Flickr

album in the world.[2] Many online news sites like CNN, Yahoo!, and BBC publish images with their stories and even provide photos which feed related to current events.

Facing such huge amounts of data, browsing and announcing photos in large-scale and heterogeneous collections is an important problem that has attracted much interest in the field of information retrieval. In order to efficiently manage them and access appropriate content, browsing and retrieval tools are required by users from various domains, including remote sensing, fashion, crime prevention, publishing, medicine, architecture, etc. For this purpose, many general image retrieval systems have been developed. There are mainly two frameworks: textual-based and content-based approaches.

For textual-based approach, many of the tools detect visual concept in images without analyzing their content, simply by matching user queries against collocated textual information. Examples include meta-data, user-annotated tag, captions, and generally text surrounding the image. In order to efficiently retrieve images, one could firstly manually annotate images that do not have tags using keywords and then carry out the search by matching their annotations with the required keywords. The most popular image search engines nowadays like Google Images and Yahoo Image use this approach. Technically, this kind of search method relies not on the image content directly, but on the textual information associated with images. However, this method quickly becomes inconceivable nowadays because tremendous amount of time and labor is required for annotating such huge amounts of data. Moreover, there exist some other problems for manual annotations like language, synonym.

In recent years, more and more attentions have been paid to machine-based visual concept detection and image classification. The visual concept is detected by their visual content, such as color, texture, shapes. It aims at detecting automatically from images high level semantic concepts, including scene Type (indoor, outdoor, landscape, etc.), objects (car, animal, person, etc.), events (travel, work, etc.), or even the sentiments (melancholic, happy, etc.), and proves to be extremely

[2] http://en.wikipedia.org/wiki/Facebook

Figure 1.1: An example of visual concept detection.

challenging because of large intra-class variations and inter-class similarities, clutter, occlusion and phase changes.

Most approaches on visual concept detection have so far focused on appropriate visual content description and have featured a dominant bag-of-visual words representation along with local visual descriptors. However, increasing works in the literature has discovered rich semantic meanings conveyed by the abundant text captions associated with images. The text associated with the image can provide a more direct gateway to image analysis and can be employed to detect image concept. Many recent works in this domain of visual concept detection propose to make joint use of user textual tags, and visual descriptions, for better bridging the gap between high level semantic concepts and low-level visual features. The work presented here is in that line and targets an effective multimodal approach for visual concept detection.

1.2 Problems and Objective

The goal of visual concept detection is to decide whether an image belongs to a certain category or not. The related PASCAL Visual Object Classes(VOC) recognition has as aim to accurately detect the bounding boxes and labels of objects in a set of images, whereas the Visual Concept Detection and Annotation(VCDA) task

focuses on both visual and textual information instead of visual information only and furthermore we offer a larger range of concepts to detect. In this thesis, different types of concepts have been considered in the literature, e.g. defined by the presence of certain concepts, such as cars or bicycles, or defined in terms of scene types, such as city, coast, mountain, indoor, outdoor, travel, etc. More precisely, only categories of objects, or concepts, are taken into account, that is to say that we want to detect any concept in an image, rather than a particular concept which is the goal of concept detection systems. An example is shown in Figure 1.1, in which the image should be classified to the predefined category "Person", "Horse", "Outdoor", "Tree", and "Building" at the same time as it contains these concepts.

The state of the art for visual concept detection using visual content focus on employing a large set of local visual features, which are extracted from a dense or sparse grid over the image and all based on the visual gradient information. Its major shortcoming is still its lack of descriptive power as regard to high level semantic concepts because of its nature of low level features. Nowadays, visual concept detection based on visual content seems to reach the performance ceiling. In order to solve this problem, the multimodal approach has been proposed, and attempt to make joint use of visual descriptions and abundant tags associated with images for better prediction of visual concepts[Wang *et al.* 2009][Guillaumin *et al.* 2010]. Despite many efforts and much progress that have been made during the past years, it remains an open problem and is still considered as one of the most challenging topics in computer vision, because it has to deal with the problems inherent to object categories, like the wide variety of shape and appearance of objects inside a category, and due to the representation of an object in an image, such as various scales and orientations, as well as illumination and occlusion problems. However, considering the texts simply interpreted as an unordered collection of words, thus disregarding grammar and even word order, the relatedness among words will not be considered. Finally, it is still a big challenge to choose the best way to integrate each modality content.

In this context, the objective of our work can be summarized as problems: (1) to propose some innovative contributions to the visual concept detection task in

particular concerning novel image feature and representation. (2) To employ the tags associated with image to build textual features and textual kernels, which are based on semantic distance. (3) To benefit from both visual models and textual models, by embedding several novel fusion approaches. These proposed approaches have been validated through experiments driven on several popular datasets.

1.3 Overview of our Approaches and Contributions

The visual concept detection task is a very challenging problem, and a lot of factors need to be considered to construct a successful system. Based on the visual content, the typical visual concept detection pipeline is composed of the following three steps:

- extraction of image features (e.g., SIFT[Lowe 2004a], DAISY[Zhu *et al.* 2011] descriptors).
- encoding of the local features in an image descriptor (e.g., a histogram of the quantized local features).
- classification of the image descriptor (e.g., by a support vector machine).

Extraction of image features aims at extracting compact and informative feature vectors or descriptors rather than using the raw data from an image to represent its visual content. The visual concept detection task depends very strongly on all the stages of the pipeline, and especially on the feature extraction step. This step plays an important role in the system, because we want that the features should be both discriminative enough and computationally efficient, while possessing some properties of robustness to changes in viewpoint, scale and lighting conditions. After features extraction, encodings for bag of visual words models have been considered. Many different encoding approaches have been proposed in the literature. Among the most successful we can cite methods such as locality-constrained linear encoding[Wang *et al.* 2010], improved Fisher encoding[Perronnin *et al.* 2010], super vector encoding[Zhou *et al.* 2010], and kernel codebook encoding[van Gemert *et al.* 2008], etc. The final step of image classification aims at constructing a robust classifier which could effectively classify images

or objects into given categories based on the extracted image feature vectors or descriptors. Many different classifiers have also been proposed in the past years, such as Support Vector Machines (SVM)[Cortes & Vapnik 1995], or Artificial Neural Networks (ANN)[Bishop 1995].

In contrast to the visual content, the text associated with images provides valuable semantic meanings about image content that can hardly be described by low-level visual features. In order to benefit from the abundant texts associated with images, the bag-of-words approach is used to organize these texts, which often is described according to the vector space model[Salton *et al.* 1975] as a vector of terms, each component of which is a kind of word count or term frequency as exemplified by tf-idf (term frequency inverse document frequency). This model has undergone several extensions, including latent semantic analysis (LSA) [Hofmann 1999a], probabilistic LSA [Hofmann 1999b] and Latent Dirichlet allocation (LDA)[Blei *et al.* 2003]. However the major drawback of these BoW-based approaches is lack of semantic sensitivity. Finally, a multimodal approach is employed to make joint use of user textual tags and visual descriptions. As multimodal approach, the gap between high level semantic concepts and low-level visual features is bridged. Meanwhile, several fusion scheme is introduced.

In this thesis, we firstly focus on image feature extraction by proposing several new image features and encoding of these novel features for the task of the visual concept detection. Meanwhile, in order to benefit from the abundant text associated with images, several novel textual descriptors are proposed to capture semantic meanings about image content. Finally the Multiple Kernels Learning(MKL) is employed to effectively fuse different features, towards automatically predicting the visual concepts of images.

1.3.1 Encoding Local Binary Descriptors by Bag-of-Features with Hamming Distance

Firstly, local binary descriptors, e.g. LBP and BRIEF[Calonder *et al.* 2010]), are becoming increasingly popular in the computer vision domain. Compared to other pop-

ular local descriptors such as SIFT, HOG, SURF and so on, binary descriptors are very fast to compute and match, as well as possessing advantages of memory and storage efficiency, because they are based directly on the binary bitstrings. They have exhibited good performances in image matching related tasks[Calonder *et al.* 2010]. However, the bottleneck of applying them in the domain of Visual Object Classes(VOC) recognition lies in the high dimensional histograms produced by encoding these binary bitstrings into decimal codes. In order to address this problem, instead of encoding the binary bitstrings into decimal codes, we propose to encode them directly by employing the BoF model with Hamming distance. The advantages are two-fold: (1) the dimensionality of the resulting histograms only depends on the size of the visual vocabulary, and is no longer related to the length of binary bitstrings, making local binary descriptors more feasible for the task of VOC recognition, especially when more bits are considered; (2) It is computationally efficient because compared to other distance measurements such as Euclidean distance, the Hamming distance is more suitable for binary descriptors, and can be computed very efficiently via a bitwise XOR operation followed by a bit count. The proposed method will be validated in the experiments section by applying on LBP feature for the purpose of VOC recognition.

The main contributions of this work considering this part are summarized as follows:

- Encoding local binary descriptors by the Bag-of-Features (BoF) model directly on binary bitstrings to address the high dimensionality issue and make them more feasible for the VOC recognition task.
- Using Hamming distance together with k-means for visual vocabulary construction and histogram assignment for computational efficiency.

1.3.2 Sampled Multi-scale Color Local Binary Patterns

Secondly, the local binary pattern (LBP) operator[Ojala *et al.* 2002a] is a computationally efficient yet powerful feature for analyzing image texture structures, and has been successfully applied to applications as diverse as texture classification, texture

segmentation, face recognition and facial expression recognition. However, the original LBP descriptor also has several drawbacks in its application. It covers a small spatial support area, hence the bit-wise comparisons are made through single circular pixel values with the central pixel. This means that the LBP codes are easily affected by noise[Liao *et al.* 2007]. Moreover, features calculated in a single circular neighborhood cannot capture larger scale structure (macrostructure) that may be dominant features. Meanwhile, the original LBP descriptor ignores all color information (its calculation is based on gray level image), while color plays an important role for distinction between objects, especially in natural scenes[Zhu *et al.* 2010]. In this work, we propose a novel representation, called Sample Multi-scale Color Local Binary Pattern (SMC-LBP), to overcome the mentioned limitations of LBP and extend the LBP feature to patch. To validate the proposed feature, we apply it to the VOC Recognition problem. In SMC-LBP, the computation is based on random sampling the neighboring pixels from multi-scale circles. Furthermore, in order to enhance the photometric invariance property and discriminative power, the proposed descriptor is computed in different color spaces. To summarize, the SMC-LBP descriptor presents several advantages:

- It encodes not only single scale, but also multiple scales of image patterns, extends the LBP to the patch, and hence provides a more complete image representation than the original LBP descriptor.
- It cooperates with color information, therefore its photometric invariance property and discriminative power are enhanced.

1.3.3 Construction of Textual descriptors

Thirdly, the tags associated with images are tended to be noisy in the sense that they are not directly related to the image content. However, there is still much information in tags. This kind of information is hard to describe by visual descriptors. Usually the term frequency model is used to represent the tags as bag-of-words (BoW), where each component of the vector is word count or term frequency. The BoW approach achieves good performance on the Visual Concept Detection and An-

notation(VCDA) task. But this approach has two main drawbacks: (1) The BoW is sensitive to the changes in vocabulary that occur when training data can not be reasonably expected to be representative of all the potential testing data; (2) The BoW only considers the word frequency information, thus disregards tags semantic information; (3) The BoW is still seriously sensitive to the changes in dictionary.

In order to solve these problems, we propose two different semantic textual features that use the textual semantic information to build the semantic features: (1) Semantic Bag-of-words(sBoW) feature, (2) Image Distance feature(IDF) based on tags associated with images. The semantic similarity between words is used in our work. Nowadays, how to estimate the semantic similarity between words is one of the longest-established tasks in natural language processing and many approaches have been developed. In this thesis, the semantic distance between tags is measured using their position in a graph such as the $WordNet$ hierarchy[Fellbaum 1998].

In the first approach, we employ $WordNet$ similarity to build textual feature during the dictionary size reduction and assignment. The main contributions of this work considering this part are summarized as follows:

- We build textual descriptors by the semantic BoW feature, in order to capture the semantic information between tags which is hardly described by the term frequency model.

- We use $WordNet$-based semantic distance for dictionary construction and histogram assignment, in order to reduce the size of the tags representation.

However, the previous approach is still sensitive to the changes in the dictionary. We expect to solve this problem by using a second approach that does not rely on dictionary contribution. The main contributions of this work considering this part are summarized as follows:

- Building Image Distance feature based on the tags associated with images. This approach can capture tags semantic information which is hardly described by the term frequencies features.

- Using *WordNet*-based semantic distance for feature construction. Two versions are proposed: The first one is based on a dictionary. The second approach is more robust because it dose not rely on a dictionary construction.

1.3.4 Visual Concept Detection and Annotation via Multiple Kernel Learning of multiple models

Finally, in order to benefit from both the visual features and the textual features, we propose here a fusion of different feature types using a multiple kernel classifier. The text associated with the image can provide a more direct gateway to image analysis and can be employed to detect image concept. Thus, the tags associated with images are used to build the frequency features and semantic features. In other hand, for visual information the VCDA task typically presents images with histograms or distribution of features from channels such as texture, color and local gradients[Siddiquie *et al.* 2009]. This means that the multiple kernel learning (MKL) approach carries out the VCDA task with a mix of the visual kernels and the textual kernels machines[Lin *et al.* 2007]. The main contributions of this work concern an effective multimodal approach for concept detection through textual descriptors and fusion with visual descriptors.

1.4 Organization of the thesis

The rest of this document is organized as follows.

- In chapter 2, a review of the main approaches and related work for the Visual Concept Detection and Annotation(VCDA) task in the literature is given. In this chapter, we introduce visual and textual models. For visual models, more attention is paid to the feature extraction and image representation(modeling). For textual models, the frequency model and Semantic model are given. Finally, the classification algorithms and fusion strategies between visual features are introduced. In addition, we introduce several standard and popular benchmarks available in computer vision community for object recognition, image

classification, Visual Concept Detection and Annotation(VCDA) tasks. Some of them will be used to carry out experiments in the following chapters.

- In chapter 3, a novel representation, called Sampled Multi-scale Color Local Binary Pattern (SMC-LBP), together with the analysis of their invariance properties is given. The experimental results on the PASCAL VOC 2007 image benchmark show significant accuracy improvement by the proposed descriptor. In addition, a novel method for encoding local binary descriptors for Visual Object Categorization (VOC) recognition is presented. Nowadays, local binary descriptors, e.g. LBP and BRIEF, have become very popular in image matching tasks because of their fast computation and matching using binary bitstrings. However, the bottleneck of applying them in the domain of VOC recognition lies in the high dimensional histograms produced by encoding these binary bitstrings into decimal codes. To solve this problem, we propose to encode local binary bitstrings directly by the Bag-of-Features (BoF) model with Hamming distance. The experimental results on the PASCAL VOC 2007 benchmark show that our approach effectively improves the recognition accuracy compared to the traditional LBP feature.

- In chapter 4, two novel methods for building textual feature definition on semantic distance are presented. Nowadays, the tags associated with images have been popularly used in the VCDA task, because they contain valuable information about image content that can hardly be described by low-level visual features. In order to solve the problem, that the term frequency model can not capture the valuable semantic information we propose the semantic bag-of-words (BoW) model which use $WordNet$-based distance to construct the codebook and assign the tags. The advantages of this approach are twofold: (1) It can capture tags semantic information that is hardly described by the term frequency model. (2) It solves the high dimensionality issue of the codebook vocabulary construction, reducing the size of the tags representation. In contrast to previous approach, we try to build semantic textual feature, called Image Distance feature(IDF) based on tags associated with im-

ages, which does not rely on a dictionary construction. The advantages of this approach are twofold: (1) It can also capture tags semantic information that is hardly described by the term frequency model. (2) It is independent of the dictionary construction, addressing feature instability.

- In chapter 5, we present a multimodal framework for Visual Concept Detection and Annotation (VCDA) task based on Multiple Kernel Learning (MKL). In the first part, we extract discriminative visual features and build visual kernels. In the second part, the tags associated with images are used to build the textual kernels. Finally, in order to benefit from both visual models and textual models, fusion is carried out by MKL efficiently embed.

- In chapter 6, our conclusions as well as some perspective for future research directions are proposed.

- In Appendix A, we present a brief description that we participate the internal contest like ImageCLEF 2011 and ImageCLEF 2012, during this thesis.

Chapter 2

Literature Review

Contents

The Visual Concept Detection and Annotation(VCDA) task is a multi-label classification challenge. The goal of this task is to decide whether a large number of images, which come from consumers, belongs to a certain concepts or not[Guillaumin *et al.* 2010]. However, the images coming from consumer include sense, events, or even sentiments. Due to large intra-class variations and inter-class

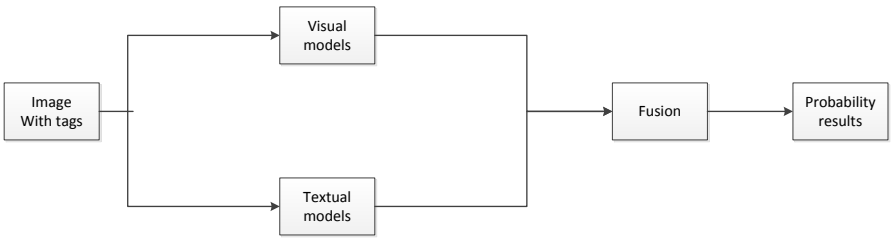


Figure 2.1: The framework of multimodel approach.

similarities, clutter, occlusion and pose changes, this work is proved to be extremely challenging in computer vision domain.

State-of-the-art methods on VCDA mostly have focused on appropriate visual content descriptors and are still less efficient on automatic textual annotation or capable of textual descriptor. Although tags associated with images from host or guest tend to be noisy in the sense that not directly relate to the image content, there is still much information in tags[Martinet *et al.* 2011]. This information is hard to describe by visual descriptor. In order to address this multi-label classification challenge task, in one hand there are good visual descriptors and textual descriptors proposed to describe the image content, in the other hand the different fusion strategies carry out the VCDA task with mix of the visual descriptor and the textual descriptor machines[Lin *et al.* 2007], as show in figure 2.1.

2.1 Introduction of Visual Models

State-of-the-art methods on Visual models are a challenging problem in computer vision. Mainly due to intra-class variations of images such as occlusion, clutter, viewpoint and lighting condition changes, these are typical in the real-world situations. In order to address these challenging problems, a lot of attention and efforts have been paid during the past decades by the researchers in computer vision community, and many approaches have been proposed in the literature. The typical pipeline includes the following three steps[Chatfield *et al.* 2011]: (1) extraction of global or local image features (e.g. SIFT[Lowe 2004b], SURF[Bay *et al.* 2008], LBP[Ojala *et al.* 2002b], etc.); (2) encoding of the local features in an image de-

scriptor (e.g. a histogram of the quantized local features), global features can be directly sent to classifiers; (3) classification of the image descriptor by certain machine learning algorithms (e.g. support vector machine, decision tree, etc.)[Chatfield *et al.* 2011]. This chapter deals with these different aspects and the approaches people have proposed for these purposes. Unfortunately, it is well known that the performance of visual models depends very strongly on all the stages of the pipeline, and especially on the feature computation step.

2.1.1 Image feature extraction

In order to build Visual Models, the direct way is the automatical extraction of feature vectors(color, texture, shape, spatial layout, etc.) using computer vision techniques. It aims at transforming the image content into a set of feature vectors(local descriptor) or a single feature vector(global descriptor). These feature vectors include a lot of redundant information and can be of very high dimension. Meanwhile, these feature vectors are also sensitive to any image variations. Many sophisticated algorithms have been designed to describe color, shape, and texture features. These algorithms are expected to adequately model image semantics and deal with broad content image. Meanwhile, the extracted features are expected to be discriminative, computationally efficient, with reasonable size, and possessed of some robustness properties to image variations(viewpoint, scale, illumination, etc.). Moreover, the following process will no longer rely on the image itself, but only on the information carried by the extracted features. Thus, feature extraction is a very important step to ensure the final good performance of visual concept detection, and can be considered as the basis of the whole process.

A lot of feature extraction methods have been proposed in the literature, and we could summarize them into two main categories: global features and local features.

2.1.1.1 Global features extraction

Global features are extracted directly from the whole image and are represented by a single vector or histogram based on the statistical analysis of the whole image, pixel

by pixel. It is expected to capture ideally the entire image content. However, this assumption is too hard to be satisfied in the reality, and the background introduces inevitably noise, particularly in the case where the object is very small compared to the size of image. Although the drawback of global features, it can still capture some useful information. In our work, we have studied and investigated the most popular ones among global features. Generally there are three categories for global features: (1) color, (2) texture and (3) shape.

- **color features**
  - *Color Histogram*[Swain & Ballard 1991]: Color histograms are the simplest and most common way for expressing the color characteristics of an image. They represent an image by modeling the color distribution of image pixels. Given a discrete color space defined by some color axes(e.g., RGB, Opponent, or HSV.), the color histogram is obtained by counting the number of times each discrete color pixels in the different image color's space which is discretized the image colors. The more number of bins are selected, the more detailed color distribution could be obtained, but the higher dimensional histogram will be generated. The number of bins is thus a trade-off between feature information and size. Color histogram is invariant to translation and rotation of the viewing axis, and robust to viewpoint change, but with no spatial information.
  - *Color Moments Vectors*[Stricker & Orengo 1995]: Color moments are defined as a very compact vector which contains the mean, variance and skewness (i.e. respectively the moments of order 1, 2 and 3 as shown in 2.1, 2.2 and 2.3) for each channel of a color space.

$$E_i = \frac{1}{N}\sum_{i=1}^{n} p_{ij} \tag{2.1}$$

$$\sigma_i = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(p_{ij} - E_i)^2} \tag{2.2}$$

$$S_i = \sqrt[3]{\frac{1}{N}\sum_{j=1}^{N}(p_{ij} - E_i)^3} \tag{2.3}$$

where i is the index of each channel, N is total number of image pixels, and p_{ij} is the value of the j-th pixel in channel i. Color moments have the same invariance properties and drawbacks as color histogram.

- *Color Coherence Vectors*[Al-Hamami & Al-Rashdan 2010]: In order to capture the spatial information of color distribution, Color Coherence Vectors are proposed. It is defined as the concatenation of two histograms, which are the population of coherent color pixels and the populations of incoherent color pixels. We say that a color is coherent when its population of pixels located in a spatial neighbor area is bigger than a predefined threshold, otherwise it is incoherent.

- *Color Correlogram and Color Auto Correlogram*[Huang *et al.* 1997] Color correlogram can be understood as a 3-dimensional matrix with size of $(n \times n \times r)$, where n is the number of color bins in an image and r is the maximal distance between two considered pixels. This matrix is indexed by color pairs, where the k-th entry for (i, j) specifies the probability of indexing a pixel of color i at a distance k away from a pixel of color j in the image. The final feature is obtained by decomposing this matrix into a single vector. As the size of color correlogram is usually too large due to its three dimensions, color auto-correlogram is also proposed to only consider the pair of pixels with the same color i at a distance k, thus resulting in a more compact representation. Their advantages are that they integrate the spatial correlation of colors and robustly tolerate large changes in appearance, viewing position and camera zoom. High computational cost is also their main drawback.

There also exist a lot of other color features in the literature, such as Dominant Color, Scalable Color, Color Layout, Color Structure, etc. These give other detail information about color. It is not possible to make an exhaustive

introduction here.

- **Texture features** Texture features is also a kind of important visual features. It can capture the content of an image efficiently. There is no precise definition for texture features. Generally, it is intuitively considered as the repeated patterns of local variation of pixel intensities, thereby quantifying the properties such as smoothness, coarseness and regularity in an image.

Table 2.1: Some texture features extracted from gray level co-occurrence matrix (GLCM)[Kurani *et al.* 2004].

Texture feature	Formula
Energy	$\sqrt{\sum_i \sum_j P_d^2(i,j)}$
Entropy	$-\sum_i \sum_j P_d^2(i,j) \mathrm{In} P_d(i,j)$
Contrast	$\sum_i \sum_j P_d(i-j)^2(i,j)$
Homogeneity	$\sum_i \sum_j \frac{P_d(i,j)}{1+(i-j)^2}$

  - *Texture Co-occurrence Matrix*[Gotlieb & Kreyszig 1990]: Gray Level Co-occurrence Matrix (GLCM) is a tabulation of how often different combinations of pixel brightness values (grey levels) occur in an image. It estimates image properties of the second order texture statistics by considering the relationship between groups of two neighboring pixels in the image. GLCM texture considers the relation between two pixels at a time, called the reference and the neighbour pixel. The neighbour pixel is chosen to be the one to the east (right) of each reference pixel. Given a displacement vector $d = (dx, dy)$, GLCM P_d of size $N \times N$ for d is calculated in such a way that the entry (i, j) of P_d is the occurrence number of the pair of gray levels i and j which are at a distance d apart. Here N denotes the number of gray levels considered in the image. Usually, the matrix P_d is not directly used in an application and a set of more compact features are computed instead from this matrix, as shown in table 2.1. The main problem of GLCM is that there is no well established method for selecting the optimal displacement vector d. In the practice, four displacement vectors are commonly used: $d = (1, 0), d = (0, 1), d = (1, 1)$

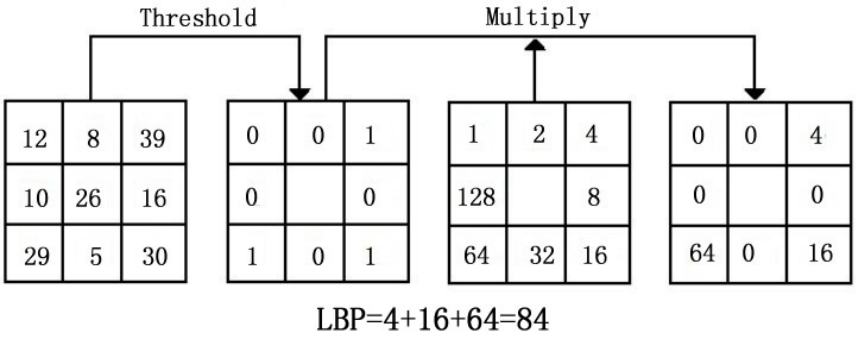


Figure 2.2: Calculation of the original LBP descriptor.

and $d = (1, 1)$.

– *Gabor*[Daugman 1988] Basically, Gabor filters are a group of wavelets, with each wavelet capturing energy at a specific frequency and a specific direction. It has been found to be particularly appropriate for texture representation and discrimination, thus the Gabor filters (or Gabor wavelets) are widely adopted for texture features extraction. Gabor filters are directly related to Gabor wavelets, since they can be designed for a number of dilations and rotations. However, in general, expansion is not applied for Gabor wavelets, since this requires computation of bi-orthogonal wavelets, which may be very time-consuming. Therefore, usually, a filter bank consisting of Gabor filters with various scales and rotations is created.

– *Local Binary Patterns* Among all these classical texture features, LBP is a more recent one and one of the most popular texture descriptors. It was introduced and used in texture classification based on local binary patterns and nonparametric discrimination of sample and prototype distributions[Ojala *et al.* 2002a]. It can be seen as a unified approach to statistical and structural texture analysis. Fig. 2.2 gives an example of LBP computation. The LBP descriptor encodes one pixel of an image by thresholding the neighborhood of each pixels with the center value. Then the threshold results are multiplied with weights given by powers of

two. Finally the LBP code is obtained by summing up all the weighted results. This process is done for each pixel, and the image representation is obtained by counting the histogram based on these codes. It creates then a global descriptor of the image. The LBP descriptor is further extended to multi-scale using a circular neighborhood with variant radius and variant number of neighboring pixels.

Because of its descriptive power for analyzing both micro and macro texture structures, and computational simplicity, LBP has been widely applied for texture classification[Ojala *et al.* 2002a] and object recognization[Zhu *et al.* 2010][Paulhac *et al.* 2008], and is demonstrated excellent results and robustness against global illumination changes. It has also been used successfully for texture segmentation[Blas *et al.* 2008][Paulhac *et al.* 2009], recognition of facial identity[Guo *et al.* 2010] and expression[Shan *et al.* 2009].

However, the original LBP descriptor also has several drawbacks in its application. It covers a small spatial support area, hence the bitwise comparisons are made through single circular pixel values with the central pixel. This means that the LBP codes are easily affected by noise[Liao *et al.* 2007]. Moreover, features calculated in a single circular neighborhood cannot capture larger scale structure (macrostructure) that may be dominant features. Meanwhile, the original LBP descriptor ignores all color information (its calculation is based on gray image), while color plays an important role for distinction between objects, especially in natural scenes[Zhu *et al.* 2010]. There can be various changes in lighting and viewing conditions in real-world scenes, leading to large variations of objects in surface illumination, scale, etc., which make the original LBP performance is not very good in Visual Concept Detection and Annotation tasks. In order to address these drawbacks, many improve method of LBP descriptors have been proposed, such as Multi-scale Block LBP[Liao *et al.* 2007], Hierarchical Multi-scale LBP[Guo *et al.* 2010], Multi-scale Color LBPs[Zhu *et al.* 2010] and so on.

- **Shape features** The shape of an object is also an important clue for recognition, especially for rigid objects. Shape is a geometrical description of the external boundary of an object, and can be described by basic geometry units such as points, lines, curves and planes. The popular shape features mainly focus on the edge or contour of an object to capture its shape information.

  - *Edge Histogram*[Swain & Ballard 1991]: Edge histogram describes edge information with a histogram based on edge distribution in an image. Five types of edges, namely vertical, horizontal, 45-degree diagonal, 135-degree diagonal and non-directional. To compute edge histogram, an image is first divided into 4×4 non-overlapping blocks, resulting in 16 equal-sized sub-images regardless of the size of the original image. In each of the sub-images, a histogram of edge distribution with 5 bins corresponding to 5 types of edges is computed, leading to a final histogram with $16 \times 5 = 80$ bins after concatenation. An extended version of edge histogram is also proposed by partitioning the image into 4×1, 1×4 *and* 2×2 sub-images in order to integrate the information of edge distribution in different scales.

  - *Line Segments*: Pujol and Chen proposed line segment based edge feature using Enhanced Fast Hough Transform (EFHT), which is a reliable and computationally efficient way of extracting line segments from an edge image. Once all the line segments are identified by EFHT, line segment based edge feature is extracted as a histogram of line segments' lengths and orientations. In order to obtain the invariant properties for scaling, translation and rotation, all the lengths are divided by the longest line segment and then an average orientation is computed so that all the angles can be expressed with respect to it. The size of the histogram is determined experimentally and set to 6 bins for orientation and 4 bins for length. Compared to the edge histogram feature, the proposed feature can provide structure information through edge connectivity while still

keeping a relatively low computational complexity.

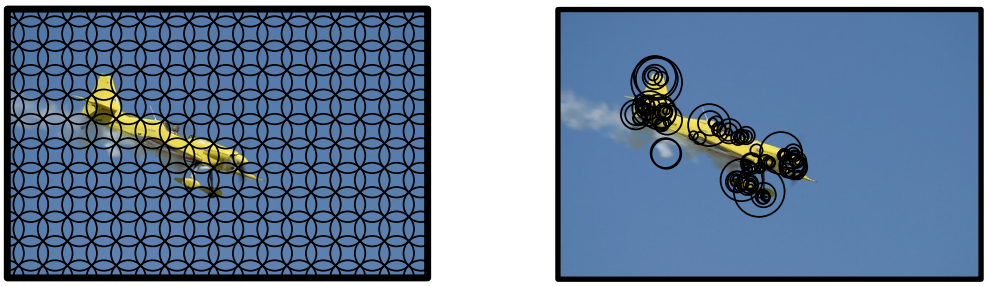
Here are some examples of possible global features. Hence again, it is impossible to give an exhaustive list.

All of global features previously introduced is in the form of a single histogram or feature vector, which is the same size for the all input images. It is not dependent to their size. Therefore, there is no requirement to transform these descriptions for a compression process. All of these global features are great sensitive to background clutter, image occlusion, and illumination variations. Moreover, these global methods implicitly assume that the objects of interest should occupy most of the region in images. However, this assumption is hard to be satisfied in real situations, where background noises always exist, particularly in the case where the object of interest is very small compared to the image size. All these limitations make global features gradually give their way to local image features.

2.1.1.2 Local features extraction

Local image features have received a lot of attention in recent years, and they have already gained the popularity and dominance in Visual Concept Detection and Annotation tasks nowadays. Compared with operating on the whole image, Local features can be points, but also edges or small image patches. Generally, the aim of local feature is to extract distinctive information which differs from its immediate neighborhood. It is usually associated with a change of an image property or several properties simultaneously, although it is not necessarily localized exactly on this change. The image properties commonly considered are intensity, color, and texture. The descriptors can then be used for various applications. By this way, local features could be more discriminative and robust to image variations, compared to the global ones. The typical local feature extraction is composed of the following two steps: (1) local keypoint/region detection and (2) local descriptor extraction.

- **Sampling Strategy** The visual appearance of a concept has a strong dependency on the viewpoint under which it is recorded. In a way, the ideal



(a) R=1,P=8 (b) R=2,P=16

Figure 2.3: Comparison of interest points/regions and dense sampling strategies for local keypoint/region detection[van de Sande *et al.* 2010].

local feature would be a point as defined in geometry: having a location in space but no spatial extent. In practice however, images are discrete with the smallest spatial unit being a pixel and discretization effects playing an important role. To localize features in images, a local neighborhood of pixels needs to be analyzed, giving all local features some implicit spatial extent. Traditionally, the term detector has been used to refer to the tool that extracts the features from the image, e.g., a corner, blob or edge detector. Salient point methods[Tuytelaars & Mikolajczyk 2007] introduce robustness against viewpoint changes by selecting points, which can be recovered under different perspectives. Another solution is to simply use many points, which is achieved by dense sampling.

- **Interest Points/Regions detector** In order to determine salient points, there is a few examples of interest points based methods. Interest points are usually keypoints located on edges or corners. Interest regions are usually regions containing a lot of information about image structures like edges and corners, or local blobs with uniform brightness. [Lindeberg 1998] developed a scale invariant blob detector, where a blob is defined by a maximum of the normalized Laplacian in scale-space. The original Harris corner detector is invariant to rotation but is not scale-invariant [Harris & Stephens 1988]. Multi-scale Harris by was adapted to solve this problem by selecting the points in the multi-

scale representation [Mikolajczyk & Schmid 2004]. The Harris-Laplace detector [Mikolajczyk & Schmid 2001] is invariant to rotation and scale changes. By applying it on multiple scales, it is possible to select the characteristic scale of a local corner using the Laplacian operator [Mikolajczyk & Schmid 2001]. Hence, for each corner the Harris-Laplace detector selects a scale-invariant point if the local image structure under a Laplacian operator has a stable maximum, as shown in figure 2.3(b).

- **Dense point detector** For concepts with many homogenous areas, like scenes, corners are often rare. Hence, for these concepts relying on a Harris-Laplace detector can be suboptimal. To counter the shortcoming of Harris-Laplace, random and dense sampling strategies have been proposed[Tuytelaars & Mikolajczyk 2007]. We employ dense sampling, which samples an image grid in a uniform fashion using a fixed pixel interval between regions, as shown in figure 2.3(a).
- **Random point detector** Other studies[Maree *et al.* 2005] have proposed to use random sampling strategy for localizing keypoints/ regions. As the name implies, keypoints/regions are randomly selected in images for local descriptor extraction. It seems that on some cases, this approach is not significantly worse than the more clever choices presented above.

Figure 2.3 shows the comparison of interest points/regions and dense sampling strategies for local keypoint/region detection. It is worth noticing that combining different strategies may provide further improvements. The winning system of the PASCAL VOC challenge 2007 demonstrated that the combination of interest points detector and dense sampling strategy performs clearly better than either of the two separately.

- **Local Visual Feature Extraction** After local keypoint/region detection, the detected regions or local neighborhood around the detected keypoints are described by local image descriptors, which should be discriminative, computationally efficient, and robust against various image variations such as scaling, affine distortions, viewpoint and illumination changes. Many different local

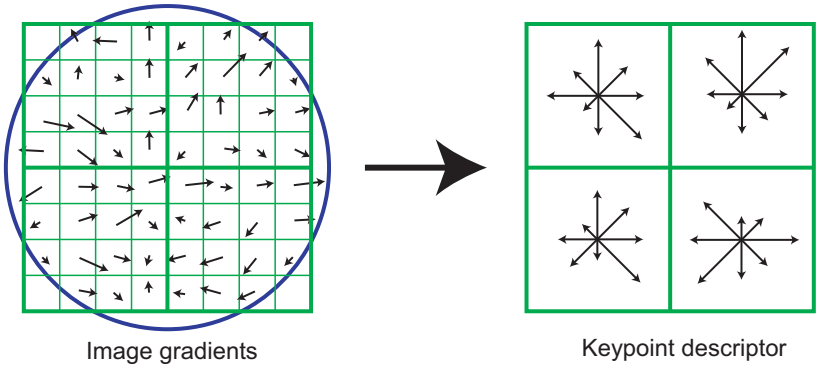


Figure 2.4: The boxes are 'plates' representing replicates. The outer plate represents text, while the inner plate represents the repeated choice of topics and words within a text[Lowe 2004a].

descriptors have been proposed in the literature, and the most popular ones are distribution-based descriptors, which represent region properties by histograms. The most popular local descriptors applied to the domain of object recognition are listed as follows:

- **Scale Invariant Feature Transform** The SIFT descriptor proposed by Lowe[Lowe 2004a][van de Sande *et al.* 2010] describes the local shape of a region using edge orientation histograms. A keypoint descriptor is created by first computing the gradient magnitude and orientation at each image sample point in a region around the keypoint location, as shown on the figure 2.4 left. These are weighted by a Gaussian window, indicated by the overlaid circle. These samples are then accumulated into orientation histograms summarizing the contents over 4×4 subregions, as shown on the figure 2.4 right, with the length of each arrow corresponding to the sum of the gradient magnitudes near that direction within the region. This figure shows a 2×2 descriptor array computed from an 8×8 set of samples, whereas the experiments in this paper use 4×4 descriptors computed from a 16×16 sample array.
- **OpponentSIFT** OpponentSIFT describes all the channels in the opponent color space using SIFT descriptors. The information in the $O3$ channel is equal to the intensity information, while the other channels

describe the color information in the image. These other channels do contain some intensity information, but due to the normalization of the SIFT descriptor they are invariant to changes in light intensity.

- **C-SIFT** In the opponent color space, the O1 and O2 channels still contain some intensity information. To add invariance to intensity changes, [Geusebroek *et al.* 2001] proposes the C-invariant which eliminates the remaining intensity information from these channels. The use of color invariants as input for SIFT was first suggested by [Abdel-Hakim & Farag 2006]. The C-SIFT descriptor[Burghouts & Geusebroek 2009] uses the C invariant, which can be intuitively seen as the normalized opponent color space O1 O3 and O2 O3. Because of the division by intensity, the scaling in the diagonal model will cancel out, making C-SIFT scale-invariant with respect to light intensity. Due to the definition of the color space, the offset does not cancel out when taking the derivative: it is not shift-invariant.

- **RGB-SIFT** For the RGB-SIFT descriptor, SIFT descriptors are computed for every RGB channel independently. An interesting property of this descriptor, is that its descriptor values are equal to the transformed color SIFT descriptor. This is explained by looking at the transformed color space: this transformation is already implicitly performed when SIFT is applied to each RGB channel independently. Because the SIFT descriptor operates on derivatives only, the subtraction of the means in the transformed color model is redudant, as this offset is already cancelled out by taking derivatives. Similarly, the division by the standard deviation is already implicitly performed by the normalization of the vector length of SIFT descriptors. Therefore, as the RGB-SIFT and transformed color SIFT descriptors are equal, we will use the RGB-SIFT name throughout this paper.

- **SURF** [Bay *et al.* 2008] proposed Speeded-Up Robust Features (SURF), which is inspired by SIFT, but several times faster to compute. Instead

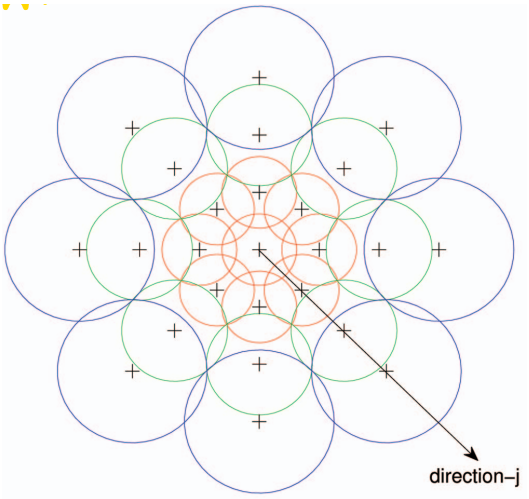


Figure 2.5: The DAISY descriptor: Each circle represents a region where the radius is proportional to the standard deviations of the Gaussian kernels and the "+" sign represents the locations where we sample the convolved orientation maps center being a pixel location where we compute the descriptor. By overlapping the regions, we achieve smooth transitions between the regions and a degree of rotational robustness. The radii of the outer regions are increased to have an equal sampling of the rotational axis, which is necessary for robustness against rotation[Tola *et al.* 2010].

of the gradient information in SIFT, SURF computes the Haar wavelet responses, and exploits integral images for computational efficiency. The input region around a keypoint is divided into 4×4 sub-regions, within which the sum of the first order Haar wavelet responses in both x and y directions are computed. The standard SURF descriptor is of 64 dimensions.

– **DAISY** [Tola *et al.* 2010] Similar to SIFT, DAISY descriptor is a 3D histogram of gradient locations and orientations. The differences between them lie in two aspects. One is that DAISY replaces the weighted sums of gradient norms used in SIFT by convolutions of gradients in specific directions with several Gaussian filters. This is for computing descriptor efficiently at every pixel location, because the histograms only need to be computed once per region and could be reused for all neighboring pixels. The other is that DAISY uses a circular neighborhood configuration instead of the rectangular one used in SIFT, as shown in figure 2.5.

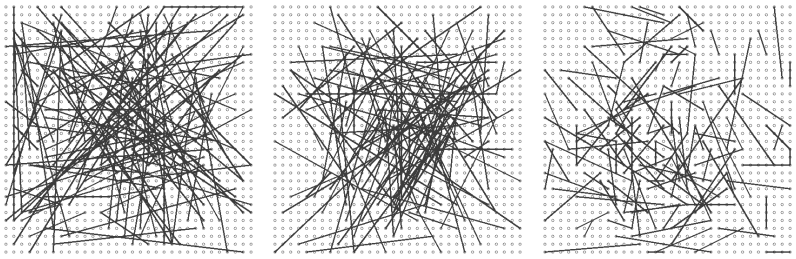


Figure 2.6: Different approaches to choosing the test locations. All except the points are selected by random sampling[Calonder *et al.* 2010].

– **HOG** [Dalal & Triggs 2005] proposed Histogram of Oriented Gradient (HOG), which is a 3D histogram of gradient locations and orientations. It is similar to both SIFT and GLOH[Mikolajczyk & Schmid 2005], because it uses both rectangular and log-polar location grids. The main difference between HOG and SIFT is that HOG is computed on a dense grid of uniformly spaced cells, with overlapping local contrast normalization. This is for better invariance to illumination and shadowing, and can be done by accumulating a measure of local histogram energy over larger spatial blocks and then using the results to normalize all of the sub-images in each block. The standard HOG descriptor is of 36 dimensions.

– **BRIEF** The BRIEF descriptor proposed by Michael Calonder[Calonder *et al.* 2010] describes the image patches pattern which could be effectively classified on the basis of a relatively small number of pairwise intensity comparisons. They propose to use binary strings as an efficient feature point descriptor, which They call BRIEF, as is shown in figure 2.6. They show that it is highly discriminative even when using relatively few bits and can be computed using simple intensity difference tests. Furthermore, the descriptor similarity can be evaluated using the Hamming distance, which is very efficient to compute, instead of the $L2$ norm as is usually done.

2.1.2 Feature Encoding Methods

After local feature extraction, each image is represented by a set of local descriptors. It is unreasonable to feed them directly into a classifier. On one hand, the dimensions of these descriptors are relatively high because of the large number of keypoints/regions (normally around thousands) in images. On the other hand, the number of local descriptors in each image varies because the number of keypoints/regions changes from one image to another. Thus, an efficient feature modelling method is required to transform these high dimensional and variable numbers of local descriptors into a more compact, informative and fixed-length representation for further classification. The baseline method is to compute a spatial histogram of visual words (quantized local features). Recent advances replace the hard quantization of features involved in this method with alternative encodings that retain more information about the original image features. This has been done in two ways: (1) by expressing features as combinations of visual words (e.g., soft quantization[van Gemert *et al.* 2008], local linear encoding[Wang *et al.* 2010]), and (2) by recording the difference between the features and the visual words (e.g., Fisher encoding[Perronnin *et al.* 2010], super-vector encoding[Zhou *et al.* 2010]).

2.1.2.1 Bag-of-Features (BoF) representation: discrete distribution

In computer vision field, the "Bag-of-Features" (BoF) model[Sivic & Zisserman 2003] [Csurka *et al.* 2004] can be applied to model an image as a discrete distribution. Compared with document model, its main idea is adapted from the "Bag-of-Words" [McCallum & Nigam 1998] (BoW) model and is to represent an image as an orderless collection of local descriptors based on an intermediate representation called "visual vocabulary". Finally, according to BoF model, the set of local descriptors is represented as a sparse vector of occurrence counts of a vocabulary of local image features. More precisely, there are two main steps: (1) visual vocabulary construction and (2) histogram encoding. The first step for the visual vocabulary construction is to convert local descriptors represented patches to "visual word", and then applies a clustering algorithm to construct

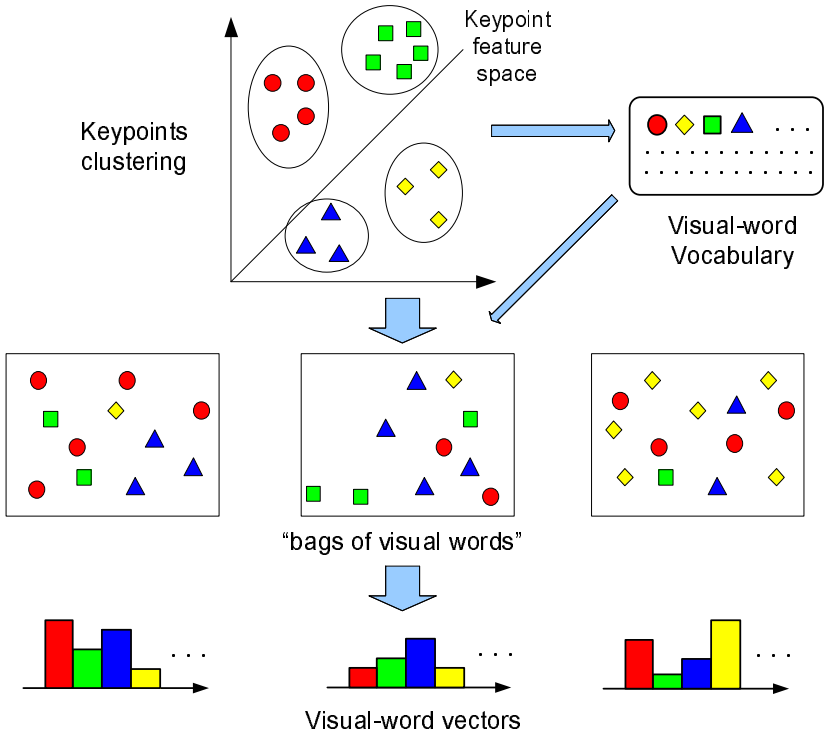


Figure 2.7: An illustration of the "Bag-of-Features" ("Bag-of-Visual-Words") method[Yang *et al.* 2007]

"visual vocabulary" on the training data. Each cluster center is considered as a "visual word" in the vocabulary. It also can be considered as a representative of several similar patches. The number of clusters is the vocabulary size. All the descriptors extracted from an image are then quantized to their closest visual word (hard assignment) or several close visual words (soft assignment) in an appropriate metric space by a certain encoding method. Thus, each local descriptor extracted from an image is mapped to a certain vocabulary through the clustering process and the assignment. An orderless collection of local descriptors extracted from an image is finally represented by a BoF vectors which is of fixed size. In other words, each image is characterized by a histogram of visual words frequencies. Figure 2.7 shows an illustration of this process. Although the BoF model is validated and shows the good performance in image related task, one of notorious disadvantages of BoW model is that it ignores the spatial relationships among the patches, which is very important in image representation. In order to address this disadvantage, researchers have proposed several methods to incorporate the spatial information.

Visual vocabulary construction There exist two methods to construct visual vocabulary offline on the training data: (1)unsupervised learning methods; (2)supervised learning methods. The $k-means$ clustering algorithm[MacQueen 1967] is the most popular one. It is unsupervised learning methods that clustering is the process of partitioning or grouping a given set of the local descriptor space into informative regions whose internal structure can be disregarded or parameterized linearly. These regions are also called visual words and a collection of visual words is called a visual vocabulary. The $k-means$ clustering method has been shown to be the most common way to construct visual vocabularies because of its relatively fast computation compared to others and more powerful methods.

The number of clusters K is assumed to be fixed in $k-means$ clustering. Given a set $x_1, \ldots, x_N \in R^D$ of N training descriptors, k-means seeks K vectors $\mu_1, \ldots, \mu_k \in R^D$ and a data-to-means assignments $q_1, \ldots, q_N \in 1, \ldots, K$ such that the cumulative approximation error $\sum_{i=1}^{N} \|x_i - \mu_{qi}\|$. The first version of this algorithm is the standard Lloyd's algorithm[Lloyd 1982], which alternates between seeking the best means given the assignments ($\mu_k = avgx_i : q_i = k$), and seek then

the best assignments given the means

$$q_{ki} = \arg min_k \|x_i - \mu_k\|^2 \tag{2.4}$$

The advantage of $k - means$ is its simple and efficient implementation, while its drawback is that most of the cluster centers are drawn irresistibly towards dense regions of the sample distribution which do not necessarily correspond to discriminative ones. In particular, the parameter k is known to be hard to choose when not given by external constraints. A radius-based clustering proposed by [Jurie & Triggs 2005], avoids setting all cluster centers into high density areas and assigns all features within a fixed radius of r to one cluster.

The unsupervised approaches is not very flexible and deficient discriminative power due to the ignorance of category information. In order to address this problem, some studies are proposed to train one special vocabulary instead of one universal vocabulary for all the training data from the whole set of categories. In [Zhang *et al.* 2007], category specific vocabularies were trained and agglomerated into a single vocabulary. Although substantial improvements were obtained, these approaches are impractical for a large number of categories as the size of the agglomerated vocabulary and the corresponding histogram representation grows linearly with the number of categories. Therefore, a compact visual vocabulary is preferred to provide a lower-dimensional representation and effectively avoid these difficulties. [Perronnin *et al.* 2006] propose to describes the content of all the considered classes of images, and class vocabularies obtained through the adaptation of the universal vocabulary using class-specific data. An image is characterized by a set of histograms - one per class - where each histogram describes whether the image content is best modeled by the universal vocabulary or the corresponding class vocabulary. Meanwhile, there are another group of methods proposed by [Yilmaz *et al.* 2008] [Liu *et al.* 2009]. All of these methods use the semantic relations between features and attempted to bring the semantic information into visual vocabulary construction.

Histogram encoding After a visual vocabulary is constructed, Histogram en-

coding is considered. This step assigns local descriptors to the visual words and characterize the visual content of an image by a histogram of visual words frequencies. Generally, there are two strategies for histogram encoding: (1) hard assignment and (2) soft assignment.

Hard assignment is a approach which employes k-means method to quantize local descriptor to a histogram. Given a vocabulary, obtained from clustering, hard assignment simply assigns the extracted local feature to their single best (usually the nearest) visual word respectively, according to a certain distance measure, as shown in equation 2.5

$$HA(\omega) = \frac{1}{N}\sum_{n=1}^{N}\begin{cases} 1 & \text{if} \quad \omega = \text{arg}min_{v \in V}(D(v, r_n)) \\ 0 & \text{otherwise} \end{cases} \tag{2.5}$$

where ω is a visual word in the vocabulary V, N is the number of local regions in an images, r_n is the local feature vector extracted from the n-th local region, and $D(v, r_n)$ is the distance between r_n and each visual word v. The hard assignment approach merely selects the best representing local features, ignoring the relevance of other candidates. Meanwhile, hard assignment plausibility denotes the problem of selecting a visual word without a suitable candidate in the vocabulary. The hard assignment approach assigns the best fitting visual word, regardless the fact that this local feature is properly representative or not [van Gemert *et al.* 2008], as illustrated in Figure 2.8.

In order to address these drawbacks of the hard assignment (a local feature is assigned to the single best visual word), there are several kinds of approaches that lead to soft assignment. The Gaussian Mixture Model (GMM) model is employed to generate the vocabulary and assign local features to each visual word by contributes to multiple visual words according to its posterior probability of the Gaussian given each visual word [J. D. H. Farquhar & Shawe-Taylor 2005][Winn *et al.* 2005][Perronnin *et al.* 2006]. Although, these works solve the word uncertainty by considering multiple visual words, they ignore visual word plausibility. In order to cope with visual word

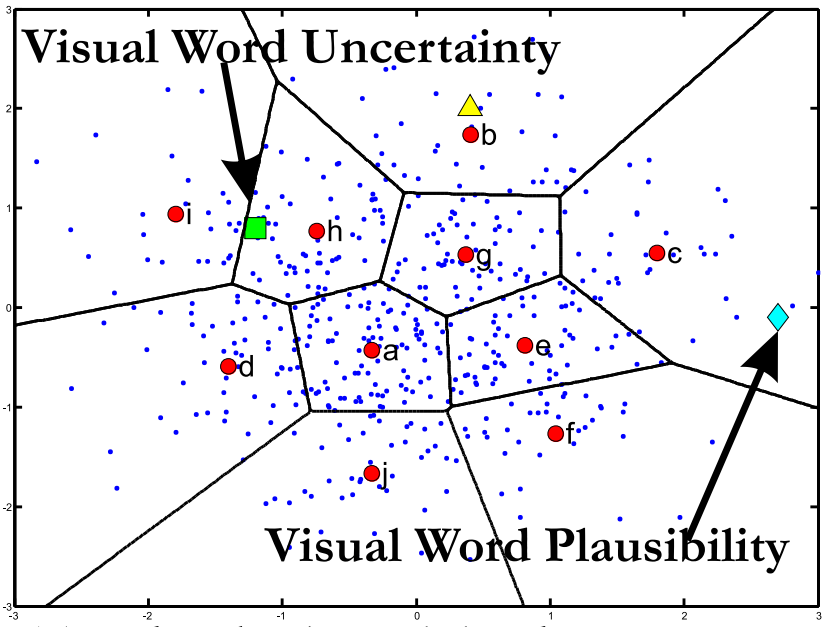


Figure 2.8: Illustration of visual word uncertainty and plausibility. The small dots represent image features, the labeled red circles are visual words found by unsupervised clustering. The triangle represents a data sample that is well suited to hard assignment approach. The difficulty with word uncertainty is shown by the square, and the problem of word plausibility is illustrated by the diamond.[van Gemert *et al.* 2008]

plausibility, [Gemert *et al.* 2008] [van Gemert *et al.* 2010] employ a decreasing function of the Euclidean distance between feature vectors and word centroids, paired with a Gaussian kernel:

$$G_\sigma(x) = \frac{1}{\sqrt{2\pi\sigma}} exp(-\frac{x^2}{\sigma^2}) \tag{2.6}$$

where σ is the smoothing parameter of kernel G. Accounted this baseline, three different formula were proposed to cope with word uncertainty (UNC), visual word plausibility (PLA) and both of them (KCB) respectively:

$$UNC(\omega) = \frac{1}{N}\sum_{n=1}^{N} \frac{G_\sigma(D(\omega, r_n))}{\sum_{k=1}^{|V|} G_\sigma(D(\upsilon_k, r_n))} \tag{2.7}$$

$$HA(\omega) = \frac{1}{N}\sum_{n=1}^{N} \begin{cases} G_\sigma(D(\omega, r_n)) & \text{if} \quad \omega = \arg min_{\upsilon \in V}(D(\upsilon, r_n)) \\ 0 & \text{otherwise} \end{cases} \tag{2.8}$$

$$KCB\sigma = \frac{1}{N}\sum_{n=1}^{N} G_\sigma(D(\omega, r_n)) \tag{2.9}$$

Recently, many new encoding methods have been proposed, such as localy-constrained linear encoding [Wang *et al.* 2010], improved Fisher encoding [Perronnin *et al.* 2010], and super vector encoding[Zhou *et al.* 2010]. All of these approaches are based on the standard histogram of quantized local features and achieved very good results on the tasks of object recognition and image classification.

Spatial information The BoF method views images as orderless distributions of local image features. All of these local image features give an equal weight, irrespective of their spatial location in the image frame. In order to overcome this limitation, [Lazebnik *et al.* 2006] proposed the "spatial pyramid" method which takes into account the spatial information of local features. They suggest to repeatedly sample fixed subregions of an image, $e.g. 1 \times 1, 2 \times 2, 4 \times 4, etc.$, and to aggregate the different resolutions into a so called spatial pyramid, which allows for region-specific

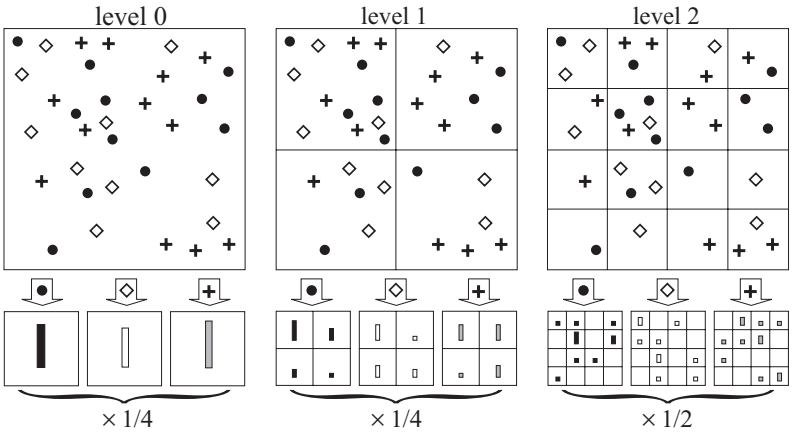


Figure 2.9: Toy example of constructing a three-level pyramid. The image has three feature types, indicated by circles, diamonds, and crosses. At the top, we subdivide the image at three different levels of resolution. Next, for each level of resolution and each channel, we count the features that fall in each spatial bin. Finally, we weight each spatial histogram.[Lazebnik *et al.* 2006]

weighting, as illustrated in Figure 2.9. Since every region is an image in itself, the spatial pyramid can be used in combination with both any kind of point detector and/or dense point sampling.

The BoF method effectively provides a mid-level representation which helps to bridge the semantic gap between low-level features extracted from an image and high-level concepts to be categorized. Its main limitation is the assumption that the distribution of feature vectors in an image can be known a priori. The optimal size of visual vocabulary, which is the basis of this approach, is also hard to be fixed.

2.1.2.2 Gaussian Mixture Model (GMM) representation: continuous distribution

Compared with discrete distribution approach, another approach was propose to model an image as a continuous distribution with the Gaussian Mixture Model (GMM). A GMM is a generative model of an input set of points where it is assumed that each point is generated independently from the same underlying probability density function (PDF). The GMM model is a weighted mixture of a set of Gaussian distributions in different parts of the input space with its own co-

variance structure[J. D. H. Farquhar & Shawe-Taylor 2005]. However the assumption is generally too restrictive and compute consumption is very huge. Therefore, [Jacob Goldberger & Greenspan 2003] [Vasconcelos *et al.* 2004] proposed to model an image as a mixture of Gaussian distributions, generally with diagonal covariance, which means that the different dimension compound to independent features.

We assume that the $x_1, \ldots, x_N \in R^D$ of N training descriptors has been generated by a parametric distribution $p(X|\lambda)$, which is considered as a Gaussian mixture model (GMM), given by

$$p(x|\lambda) = \sum_{i=1}^{M} \omega_i g(x|\mu_i, \Sigma_i). \tag{2.10}$$

$$g(x|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}}. \tag{2.11}$$

where $\lambda = (\omega_i, \mu_i, \Sigma_i)$ is the vector of parameters of the model, including the prior probability of mixing coefficients $\omega_i \in R_+$, $\sum_{i=1}^{M} \omega_i = 1$, the means value $\mu_i \in R^D$, and the positive definite covariance matrices $\Sigma_i \in R^{D\times D}$ of each Gaussian component. Here the covariance matrices are assumed to be diagonal, so that the GMM is fully specified by $(2D+1)K$ scalar parameters. In order to estimate GMM model parameters, the expectation maximization algorithm(EM) is employed from a training set of descriptors $x_1, \ldots, x_N \in R^D$ of N.

Moreover, a GMM adapted from a common "universal" GMM using the maximum a posteriori (MAP) criterion is proposed by [Liu & Perronnin 2008]. MAP provides a more accurate estimate of the GMM parameters compared to standard maximum likelihood estimation (MLE) in the challenging case where the cardinality of the vector set is small. Meanwhile, there is a correspondence between the Gaussians of two GMMs adapted from a common distribution and one can take advantage of this fact to compute efficiently the probabilistic similarity.

Those approaches model a local feature set with a continuous distribution. The most commonly used measures of similarity between two GMMs are the Kullback-Leibler divergence (KLD) [Jacob Goldberger & Greenspan 2003] [Vasconcelos 2004] and the probability product kernel (PPK) [Jebara & Kondor 2003]

[Jebara Tony & Andrew 2004].

The GMM method has two main shortcomings. Firstly, the robust estimation of the GMM parameters may be difficult as the cardinality of the vector set is small. Secondly, it is expensive to compute the similarity between two GMMs. Therefore, we choose the BoF method for image modelling in our work presented in the following chapters.

2.2 Introduction of Textual Models

The text associated with images provides valuable information about image content that can hardly be described by low-level visual features. Examples include meta-data (e.g., the image's file name and format), user-annotated tags, captions, and generally text surrounding the image. In order to capture this useful information to assist the visual concept detection tasks, a lot of textual features have been proposed in the literature, and we could summarize them into two main categories: frequency textual features and semantic textual features.

2.2.1 Preprocessing

The first step in text categorisation is to transform documents, which typically are strings of characters, into a representation suitable for the learning algorithm and the classification task. The text transformation usually is of the following kind:

- Remove stopwords
- Perform word stemming

The stopwords are frequent words that carry no information (i.e. pronouns, prepositions, conjunctions etc.). By word stemming we mean the process of suffix removal to generate word stems. This is done to group words that have the same conceptual meaning, such as walk, walker, walked and walking. The Porter stemmer is a well-known algorithm for this task.

2.2.2 Frequency textual feature

Perhaps the most commonly used methods for text associated with the image representation is the so called vector space model[Salton & Mcgill 1986]. In the vector space model, texts are represented by vectors of words. Usually, a collection of texts is represented by a word-by-text matrix A, where each entry represents the occurrences of a tag in a text, i.e.,

$$A = a_{ik} \tag{2.12}$$

where a_{ik} is the weight of tags i in text k. Since every tag does not normally appear in each text, the matrix A is usually sparse. The number of rows of the matrix M corresponds to the number of words in the dictionary. Usually, M is very large. Hence, a major characteristic or difficulty of building textual features are the high dimensionality of the feature space. For frequency textual features, the most of the approaches are based on two empirical observations regarding words.

- The more times a word occurs in a document, the more relevant it is to the topic of the document.
- The more times the word occurs throughout all documents in the collection. the more poorly it discriminates between documents.

Let f_{ik} be the frequency of word i in document k, N the number of documents in the collection. M the number of words in the collection after stopword removal and word stemming, and n_i the total number of document in the collection for which the word i at last once occurs. In what follows we describe several different weighting schemes that are based on these quantities.

2.2.2.1 Word frequency weighting

A simple approach is to use the frequency of the tags in the text:

$$a_{ik} = f_{ik} \tag{2.13}$$

2.2.2.2 tf/idf-weighting

The previous schemes do not take into account the frequency of the word throughout all texts in the collection. In order to solve this problem, a well-known approach tf/idf-weighting[Jones 1972] for computing word weights was proposed, which assigns the weight to tag i in text k in proportion to the number of occurrences of the word in the text, and in inverse proportion to the number of documents in the collection for which the word occurs at least once.

$$a_{ik} = f_{ik} * log(\frac{N}{n_i}) \tag{2.14}$$

2.2.2.3 tfc-weighting

Usually the document/text in the collection have different lengths. The tf/idf-weighting does not take into account the document/text lengths. In order to add this information, the tfc-weighting[Salton & Buckley 1988] was proposed. It considers that length normalisation is used as part of the word weighting formula.

$$a_{ik} = \frac{f_{ik} * log(\frac{N}{n_i})}{\sqrt{\sum_{j=1}^{M}[f_{jk} * log(\frac{N}{n_j})]^2}} \tag{2.15}$$

2.2.2.4 Entropy weighting

In information theory, entropy is a measure of the uncertainty in a random variable. It can quantify the expected value of the information contained in a message. Based on the sophisticated entropy theoretic idea, Entropy-weighting[Dumais 1991] was proposed. In the entropy-weighting scheme, the weight for word i in document/text k is given by:

$$a_{ik} = log(f_{ik} + 1) * (1 + \frac{1}{log(N)} \sum_{j=1}^{N}[\frac{f_{ij}}{n_i} log(\frac{f_{ij}}{n_i})]) \tag{2.16}$$

where

$$(1 + \frac{1}{log(N)} \sum_{j=1}^{N} [\frac{f_{ij}}{n_i} log(\frac{f_{ij}}{n_i})]) \tag{2.17}$$

is the average uncertainty or entropy of word i. This quantity is -1 if the word is equally distributed over all documents/texts, and 0 if the word occurs in only one document/text.

2.2.3 Semantic textual feature

The previous approches based on "bag-of-words" approach fall short to describe the fineness and the relatedness of semantic concepts. Indeed, these BoW kind approaches assume that word terms are basically statistically independent, thereby mismatching text documents close in content but with different term vocabulary. A solution to limit this problem is to detect the connection and associate words of the same meaning or words that rely on the same concepts. This can be done by the use of WordNet.

2.2.3.1 WordNet

WordNet was created at the Cognitive Science Laboratory of Princeton University under the direction of psychology professor George A. Miller.[1] It is a lexical database for the English language[Miller *et al.* 1990]. It groups English words into sets of synonyms called synsets, provides short, general definitions, and records the various semantic relations between these synonym sets. The purpose is twofold: to produce a combination of dictionary and thesaurus that is more intuitively usable, and to support automatic text analysis and artificial intelligence applications.

WordNet distinguishes between nouns, verbs, adjectives and adverbs because they follow different grammatical rules. It does not include prepositions, determiners etc. Every synset contains a group of synonymous words or collocations (a collocation is a sequence of words that go together to form a specific meaning, such as 'car pool'); different senses of a word are in different synsets. The meaning of the synsets is further clarified with short defining glosses. While semantic relations

[1] http://wordnet.princeton.edu/

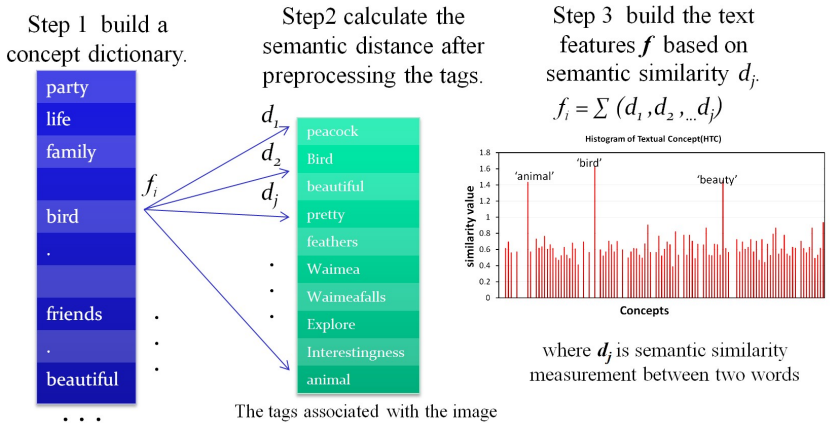


Figure 2.10: The three steps process of HTC algorithm[Liu *et al.* 2013].

apply to all members of a synset because they share a meaning but are all mutually synonyms, words can also be connected to other words through lexical relations, including antonyms (opposites of each other) which are derivationally related, as well.

Based on the structure and content of WordNet, measure of similarity use information found in a same hierarchy of concepts (or synsets), and quantify how much concept A is like (or is similar to) concept B. For example, such a measure might show that an automobile is more like a *boat* than it is a *tree*, due to the fact that *automobile* and *boat* share *vehicle* as an ancestor in the WordNet noun hierarchy[Pedersen *et al.* 2004]. A lot of measure approaches of similarity distance based on WordNet have been proposed in the literature, such as $Hirst - St - Onge$[Hirst & St-Onge 1998], $Leacock - Chodorow$[Leacock & Chodorow 1998], $Resnik$[Rubenstein & Goodenough 1965], Lin[Lin 1998] and so on.

2.2.3.2 HTC: a Histogram of Textual Concepts

In contrast with the classical bag-of-words approach which simply relies on term frequencies, Liu proposes a new textual descriptor, namely Histogram of Textual Concepts (HTC), which accounts for the relatedness of semantic concepts in accumulating the contributions of word terms toward a dictionary[Liu *et al.* 2013].

This Histograms of Textual Concepts (HTC) captures the semantic relatedness of concepts. HTC is inspired from a model that we can call componential space model, such as conceptual vector[Schwab *et al.* 2002], which describes the meaning of a word by its atoms, its components, attributes, behavior, related ideas, etc. For instance, the concept of "rain" can be described by water, liquid, precipitation, dripping liquid, monsoon, etc., thus in much a similar way when users tag photos. Similarly, the concept "peacock", as illustrated in Figure 2.10 can be described by bird, male, beautiful, pretty, feathers, plumage, animal, etc.

Specifically, the HTC of a text document is defined as a histogram of textual concepts toward a vocabulary or dictionary where each bin of this histogram represents a concept of the dictionary, whereas its value is the accumulation of the contribution of each word within the text document toward the underlying concept according to a predefined semantic similarity measure. Given a dictionary D and a semantic similarity measurement S between any pair of two terms, HTC can be simply extracted from the tags of an image through a three steps process. Note that tags such as $peacock, bird, feathers, animal$ all contribute to the bin values associated to the *animal* and *bird* concepts according to a semantic similarity measurement whereas tags such $beautiful, pretty, interestingness$ all help peak the bin value associated to the concept $beautiful$. This is in clear contrast to the BoW approaches where the relatedness of textual concepts is simply ignored as word terms are statistically counted.

2.2.4 Dimensionality reduction

The central problem of the domain remains the high dimensionality of the feature space in statistical text classification. For each unique word found in the collection of texts there exists one dimension. This means that the dimensionality is typically hundreds of thousands. Standard classification techniques can not deal with such a large feature set, since processing is extremely costly in computational terms, and the results become unreliable due to the lack of sufficient training data. Hence, there is a need for a reduction of the original set, which is commonly known as dimensionality reduction in the pattern recognition literature. Most of the dimensionality

reduction approach can be classified into one of two categories: feature selection or re-parameterisation.

2.2.4.1 Feature Selection

Feature selection is the process of selecting a subset of relevant features for use in model construction. The central assumption when using a feature selection technique is that the data contains many redundant or irrelevant features. Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful information in any context. Feature selection attempts to remove non-informative ones and creates new features from functions of the original features, whereas feature selection returns a subset of the features. As feature selection algorithm, there are three main benefits: (1)categorisation effectiveness; (2)reduce computational complexity; (3)enhanced generalisation by reducing overfitting. A lot of feature selection algorithms have been proposed in the literature, such as Document Frequency Thresholding[Yang & Pedersen 1997], χ^2-statistic[Zheng 2004], Information gain[Mori 2002], and so on.

2.2.4.2 Re-parameterisation

Unfortunately the feature selection approaches provide a relatively small amount of reduction in description length and reveals little in the way of inter- or intra-document statistical structure. In order to address these shortcoming, researchers have proposed several other dimensionality reduction techniques to find short descriptions of the members of a collection that enable efficient processing of large collections while preserving the essential statistical relationships that are useful for basic tasks such as classification. There are many generative probabilistic models for collections of discrete data such as text corpora. In this section we describe Latent Semantic Analysis(LSA)[Deerwester *et al.* 1990], probabilistic Latent Semantic Analysis (pLSA)[Hofmann 1999a], and Latent Dirichlet Allocation(LDA)[Blei *et al.* 2003].

- **Latent Semantic Analysis** LSA is based on the assumption that there is

some underlying or latent structure in the pattern of word usage across documents, and that statistical techniques can be used to estimate this structure. It uses a singular value decomposition(SVD) of the word-by-document A matrix to identify a linear subspace in the space of $tf-idf$ features that captures most of the variance in the collection. Assuming that A is $M \times N$ matrix, where M is the number of words, and N the number of documents, the singular value decomposition of A is given by:

$$A = U\Sigma V^T \tag{2.18}$$

where U and V have orthonormal columns and Σ is the diagonal matrix of singular values. The rank of A is R. If the singular values of Σ are ordered by size, the K largest may be kept and the remaining smaller ones set to zero. The product of the resulting matrices is a matrix A_k which is an approximation of A at rank K

$$A_K = U_K \Sigma_K V_K^T \tag{2.19}$$

where Σ_k is obtained by deleting the zero rows and columns of Σ, and U_K and V_K are obtained by deleting the corresponding rows and columns of U and V. This approach can achieve significant compression in large collections. Furthermore, Deerwester et al. argue that the derived features of LSI, which are linear combinations of the original $tf - idf$ features, can capture some aspects of basic linguistic notions such as synonymy and polysemy[Deerwester *et al.* 1990].

- **Probabilistic Latent Semantic Analysis** Compared to standard Latent Semantic Analysis which stems from linear algebra and performs a Singular Value Decomposition of co-occurrence tables, the PLSA is based on a mixture decomposition derived from a latent class model. This results in a more principled approach which has a solid foundation in statistics. In order to avoid overfitting, Thomas Hofmann propose a widely applicable generalization of maximum likelihood model fitting by tempered EM[Hofmann 1999a]. mod-

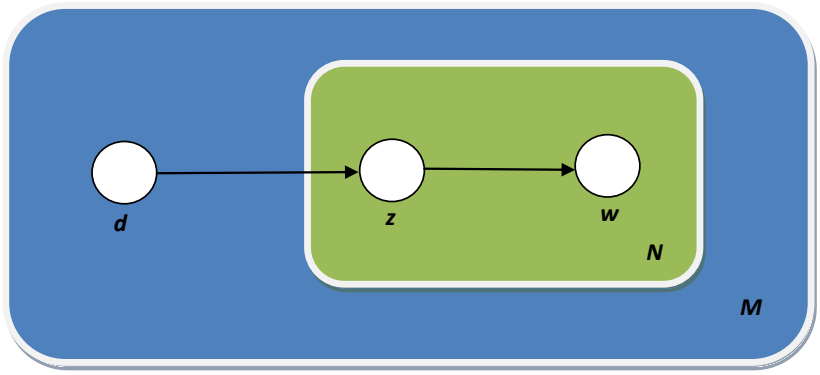


Figure 2.11: The boxes are 'plates' representing replicates. The outer plate represents text, while the inner plate represents the repeated choice of topics and words within a text.

els each word in a document as a sample from a mixture model, where the mixture components are multinomial random variables that can be viewed as representations of topics. Thus each word is generated from a single topic, and different words in a document may be generated from different topics. Each document is represented as a list of mixing proportions for these mixture components and thereby reduced to a probability distribution on a fixed set of topics. This distribution is the reduced description associated with the document. While Hofmann's work is a useful step toward probabilistic modeling of text, it is incomplete in that it provides no probabilistic model at the level of documents. In pLSI, each document is represented as a list of numbers (the mixing proportions for topics), and there is no generative probabilistic model for these numbers. This leads to several problems: (1) the number of parameters in the model grows linearly with the size of the corpus, which leads to serious problems with overfitting, and (2) it is not clear how to assign probability to a document outside of the training set. The pLSI model, illustrated in Figure 2.11, posits that a document label d and a word w_n are conditionally dependent given an unobserved topic z:

$$p(d, w_n) = p(d) \sum_z p(w_n|z)p(z|d). \tag{2.20}$$

The pLSI model attempts to relax the simplifying assumption made in the mixture of unigrams model that each document is generated from only one topic. In a sense, it does capture the possibility that a document may contain multiple topics since $p(z|d)$ serves as the mixture weights of the topics for a particular document d. However, it is important to note that d is a dummy index into the list of documents in the training set. Thus, d is a multinomial random variable with as many possible values as there are training documents and the model learns the topic mixtures $p(z|d)$ only for those documents on which it is trained. For this reason, pLSI is not a well-defined generative model of documents. There is no natural way to use it to assign probability to a previously unseen document. A further difficulty with pLSI, which also stems from the use of a distribution indexed by training documents, is that the number of parameters which must be estimated grows linearly with the number of training documents. The parameters for a k-topic pLSI model are k multinomial distributions of size V and M mixtures over the k hidden topics. This gives $kV + kM$ parameters and therefore linear growth in M. The linear growth in parameters suggests that the model is prone to overfitting and, empirically, overfitting is indeed a serious problem.

- **Latent Dirichlet Allocation** The basic idea of LDA is that documents are represented as random mixtures over latent topics. The LDA is a topic model. We expect that this topic model can help us get some underlying or latent structure in the pattern of word usage across the text and reduce the high dimensionality of the feature space. As we have hoped, these distributions seem to capture some of the underlying topics in the corpus. where each topic is characterized by a dirichlet distribution over words, in which the dimensionality k of the distribution (and thus the dimensionality of the topic variable z) is assumed known and fixed.

  LDA assumes the following generative process for each word w in a corpus D [Blei *et al.* 2003]:

  – 1. Choose $N \sim$Poisson(ξ).

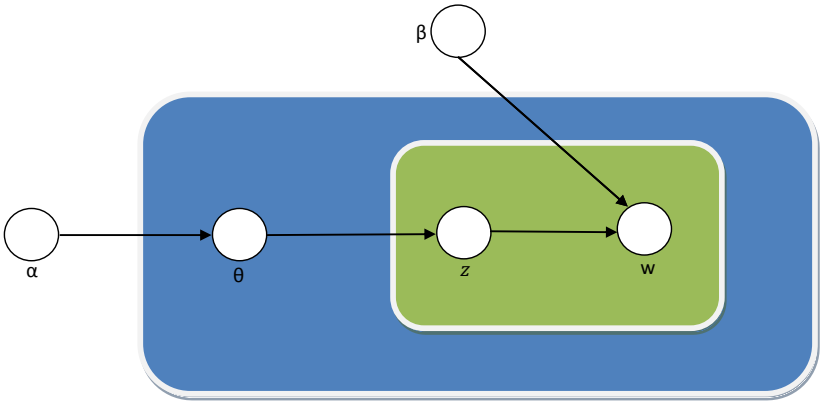


Figure 2.12: The boxes are 'plates' representing replicates. The outer plate represents text, while the inner plate represents the repeated choice of topics and words within a text.

- 2. Choose θ ∼Dirichlet(α).
- 3. For each of the N words w_n do
  - (a) Choose a topic $z_n \sim$ Multinomial(θ).
  - (b) Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n.
  - end

Graphical model representation of LDA:

There are three levels for the LDA representation. The parameters α and β are corpus-level parameters, assumed to be sampled once in the process of generating a corpus. The variables θ are document-level variables, sampled once per document. Finally, the variables z and w are word-level variables and are sampled once for each word in each document. α is the parameter of the uniform Dirichlet prior on the per-document topic distributions. β is the parameter of a $k \times V$ matrix where $\beta_i j = p(w^j = 1|z^i = 1)$. K is the dimensionality of the Dirichlet distribution(and thus the dimensionality of the topic variable z). V is the size of the dictionary.

Given α and β, the joint distribution of a topic mixture θ, a set of N topics

z, and a set of N words w is given by:

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha)\Pi_(n = 1)^N p(z_n|\theta)p(w_n|z_n, \beta) \tag{2.21}$$

The EM algorithm are used to find the Dirichlet parameter α and conditional multinomial parameter β.

We describe latent Dirichlet allocation (LDA), as a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document. We present efficient approximate inference techniques based on variational methods and an EM algorithm for empirical Bayes parameter estimation. We report results in document modeling, text classification, and collaborative filtering, comparing to a mixture of unigrams model and the probabilistic LSA model.

2.3 Classification

Based on image representations computed from the extracted features, certain pattern recognition algorithms (classifiers) are required to perform the final classification. There exist two main kinds of approaches in the literature for making the final classification: (1) generative methods and (2) discriminative methods. Generative methods produce a probability density model over all the variables and then adopt it to compute classification functions. Differently, discriminative methods directly estimate the posterior probabilities for classification without attempting to model the underlying probability distributions.

2.3.1 Generative methods

Suppose the x is the set of features representing an image to be classified, and $C_m, m = 1, \ldots, M$ being a set of class labels in consideration, the generative model will estimate the posterior probability $p(C_m|x)$ in the probabilistic framework, according to which x will be classified into the target class (for instance, if we wish to minimize the number of misclassifications, we assign x to the class having the largest posterior probability). In the case of discriminative models, the objective is to learn the precise boundaries between the different classes of samples in a multi-dimensional space (often the feature space) so that the classification can be performed by considering the position of the image projection in this space.

According to the Bayes theorem, the posterior probability $p(C_m|x)$ can be expressed in the following form:

$$p(C_m|x) = \frac{p(x|C_m)p(C_m)}{p(x)}. \tag{2.22}$$

where $p(C_m)$ is the prior probability of the class C_m, $p(x|C_m)$ is probability density of class C_m, called likelihood. $p(x)$ is the probability density over all the classes. As it is constant when considering the posterior probability for each class, its computation is not necessary. Moreover, if we know that the prior probabilities are equal, or if we make this assumption, the decision can be realized only depending on the likelihood function $p(x|C_m)$ for each class.

The typical generative method relies on a GMM to model the distribution of the training samples. A mixture model is a probabilistic model for representing the presence of subpopulations within an overall population, without requiring that an observed data set should identify the sub-population to which an individual observation belongs. Formally a mixture model corresponds to the mixture distribution that represents the probability distribution of observations in the overall population. However, while problems associated with "mixture distributions" relate to deriving the properties of the overall population from those of the sub-populations, "mixture models" are used to make statistical inferences about the properties of the sub-populations given only observations on the pooled population, without sub-

population identity information.

A GMM is a weighted sum of M component Gaussian densities as given by the equation,

$$p(x|\lambda) = \Sigma_{i=1}^{M} \omega_i g(x|\mu_i, \Sigma_i). \tag{2.23}$$

$$g(x|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}}. \tag{2.24}$$

where ω_i are the mixture weights, and $g(x|\mu_i, \Sigma_i)$ are the component Gaussian densities. Each component density is a D-variate Gaussian function of the form, with mean vector μ_i and covariance matrix Σ_i. The mixture weights satisfy the constraint that $\Sigma_{i=1}^{M} \omega_i = 1$. The complete Gaussian mixture model is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. These parameters are collectively represented by the notation $\lambda = \omega_i, \mu_i, \Sigma_i,\ i = 1, \ldots, M$. Given training vectors and a GMM configuration, the parameters of the GMM, λ, is estimated. We try to make, in some sense best, the best match with the distribution of the training feature vectors. There are several techniques available for estimating the parameters of a GMM . By far the most popular and well-established method is maximum likelihood (ML) estimation.

The aim of ML estimation is to find the model parameters which maximize the likelihood of the GMM given the training data. For a sequence of T training vectors $X = x_1, \ldots, x_T$, the GMM likelihood, assuming independence between the vectors can be written as,

$$In(p(X|\lambda)) = In(\prod_{t=1}^{T} p(x_t|\mu, \Sigma, \pi)) = In \prod_{t=1}^{T} \{\Sigma_{k=1}^{M} \omega_k g(x_t|\mu_k, \Sigma_k)\} \tag{2.25}$$

Then, we can employ the EM algorithm to maximize this likelihood function for the class C_m with respect to the parameters of the GMM. The basic idea of the EM algorithm is, beginning with an initial model λ, to estimate a new model $\bar{\lambda}$, such that $p(X|\bar{\lambda}) \geq p(X|\lambda)$. The new model then becomes the initial model for the next iteration and the process is repeated until some convergence threshold is reached.

On each EM iteration, the following re-estimation formulas are used which guar-

antee a monotonic increase in the model's likelihood value,

Mixture Weights

$$\bar{\omega_i} = \frac{1}{T}\sum_{t=1}^{T} Pr(i|x_t, \lambda). \tag{2.26}$$

Means

$$\bar{\mu_i} = \frac{\sum_{t=1}^{T} Pr(i|x_t, \lambda)x_t}{\sum_{t=1}^{T} Pr(i|x_t, \lambda)}. \tag{2.27}$$

Variances (diagonal covariance)

$$\bar{\mu_i} = \frac{\sum_{t=1}^{T} Pr(i|x_t, \lambda)x_t^2}{\sum_{t=1}^{T} Pr(i|x_t, \lambda)} - \bar{\mu}^2. \tag{2.28}$$

where $\bar{\sigma}_i^2, \bar{x}_t$, and $\bar{\mu}_i$ refer to arbitrary elements of the vectors σ_i^2, x_t, and μ_i, respectively. The a posteriori probability for component i is given by

$$Pr(i|x_t, \lambda) = \frac{\omega_i g(x_t|\mu_i, \Sigma_i)}{\sum_{k=1}^{M} \omega_k g(x_t|\mu_k, \Sigma_k)}. \tag{2.29}$$

After the optimized GMMs for all the classes are obtained, each new sample will be assigned to the class with the maximum value of the logarithm of the likelihood function.

Generative methods offer the advantage of easily adding new classes or new data for a certain class by training the model only for the concerned class rather than for all the classes. It can also deal with the situation of incomplete data. Its main drawback lies in high computational cost of learning process.

2.3.2 Discriminative methods

The objective of discriminative methods is to learn the precise boundaries between different classes of samples in a multi-dimensional space (usually the feature space) so that the classification can be performed by considering the position of the image projection in this space. Many discriminative classifiers are reported in the literature, and the kernel-based ones are the most popular.

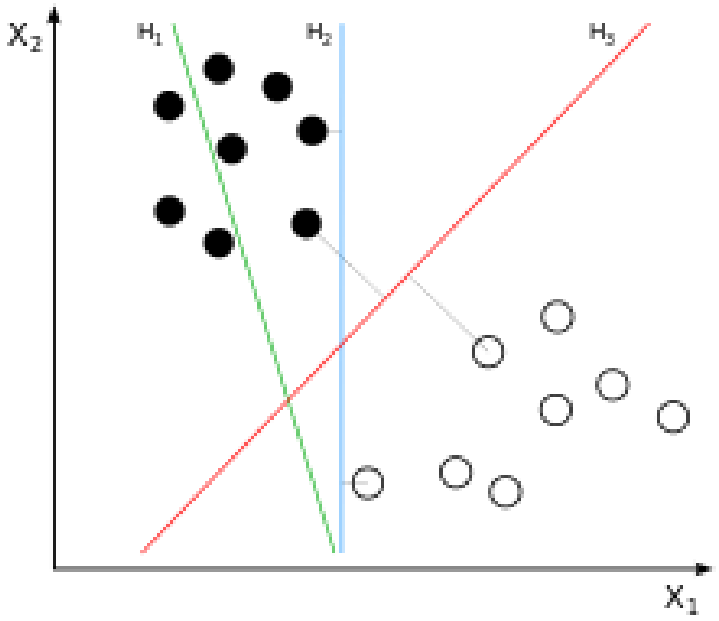


Figure 2.13: H_1 does not separate the classes. H_2 does, but only with a small margin. H_3 separates them with the maximum margin.

2.3.2.1 Support vector machine

The Support Vector Machine(SVM) is supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. It is proposed by Corinna Cortes[Cortes & Vapnik 1995] based on his statistical learning theory. The SVM constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier. An example of good separation hyperplane is illustrated in Figure 2.13. New samples are then mapped into the same space and predicted to a class, based on which side of the hyperplane they fall into.

- **Linear SVM** The standard SVM is a linear classifier for binary classification problem. Assume that we have M training samples, where each x_i has D dimensionality and is in one of two classes $y_i = -1$ or $+1$, training data is the

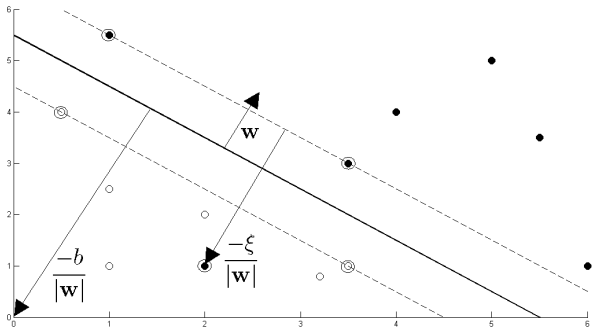


Figure 2.14: An illustration of maximum-margin hyperplane for an SVM trained with samples from two classes (samples on the margins are called the support vectors)[Fletcher 2008].

form:

$$\{x_i, y_i\} \quad where \;\; i = 1, \ldots, M, \quad y_i \in \{-1, 1\}, x \in \mathbb{R}^D. \tag{2.30}$$

Here the data is assumed linearly separable, meaning that SVM can construct a $(D-1)$-dimensional hyperplane with the maximum margin in the feature space to linearly separate these samples into two predefined classes, as illustrated in Figure 2.14. This hyperplane can be described by $w * x + b = 0$ where:

- w is normal to the hyperplane.
- $\frac{b}{\|w\|}$ is the perpendicular distance from the hyperplane to the origin.

By using geometry, we therefore need to solve optimization problem:

$$\begin{aligned} &min\{\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{L}\xi_i\}. \\ s.t. \quad &y_i(x_i \cdot w + b) - 1 + \xi_i \geq 0 \quad \forall i \quad \xi_i \geq 0 \end{aligned} \tag{2.31}$$

where the parameter C controls the trade-off between the slack variable penalty and the size of the margin. Reformulating as a Lagrangian, which as before we need to minimize with respect to $w, b \; and \; \xi_i$.

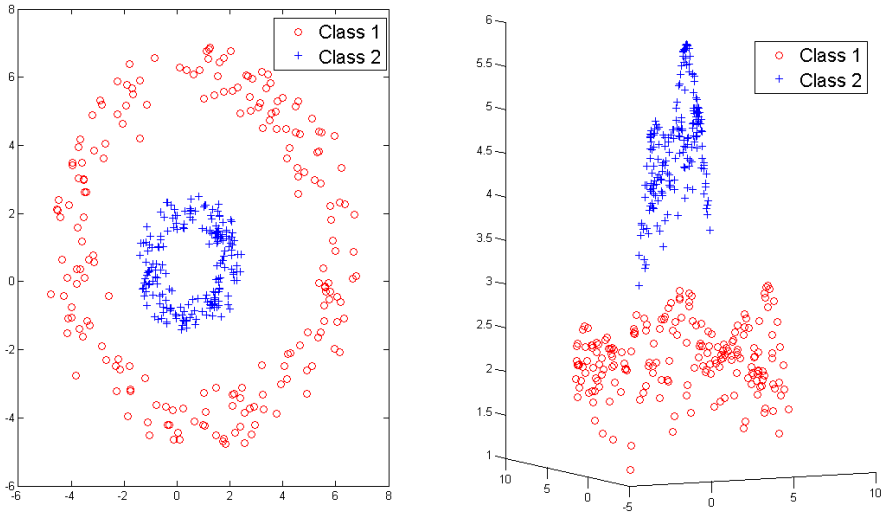


Figure 2.15: Dichotomous data re-mapped using Radial Basis Kernel[Fletcher 2008].

For a new sample x to be classified, the final decision function is in the form:

$$f(x) = sgn\{\sum_{i=1}^{N} \alpha_i^* y_i (x_i * x) + b^*\} \tag{2.32}$$

where α_i^* and b^* are the optimized parameters obtained in the training process.

- **Non-linear SVM** The original optimal hyperplane algorithm was proposed to address linearly separable problem. However, it often happens that the samples to be classified are not linearly separable in the original space. In order to overcome this drawback, in 1992, Bernhard E. Boser, Isabelle M. Guyon and Vladimir N. Vapnik proposed a way to create nonlinear classifiers by applying the kernel trick to maximum-margin hyperplane[Boser *et al.* 1992]. This approach map the samples from the original finite dimensional space into a higher or infinite dimensional space, in which these samples are supposed to be linear and the separation of them is much easier than in the original space, as shown 2.15. The family of functions which transform the original input space is called Kernel Functions $K(*,*)$.

  For the training of the non-linear SVM classifier, the equation of optimization

problem in the linear SVM training is changed as:

$$\begin{aligned} & min\{\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{L}\xi_i\}. \\ s.t. \quad & y_i(\phi(x_i)\cdot w + b) - 1 + \xi_i \geq 0 \quad \forall i \quad \xi_i \geq 0 \end{aligned} \tag{2.33}$$

where the training sample x_i are mapped into a higher or infinite dimensional space by the mapping function ϕ.

Finally, The final decision function for a new sample x is thus changed as:

$$f(x) = sgn\{\sum_{i=1}^{N}\alpha_i^* y_i K(x_i, x) + b^*\}. \tag{2.34}$$

where

$$K(x_i, x) = \phi(x_i)^T\phi(x). \tag{2.35}$$

In the final decision function, the function 2.35 and the tuning of its parameters play an important role for the non-linear SVM to achieve a good classification performance. However there is no specific application to select the kernel and the tuning of its parameters. Until now it is done empirically and experimentally, or by cross-validation in some cases.

- **Multi-class SVM** Multi-class SVM aims to assign labels to instances by using support vector machines, where the labels are drawn from a finite set of several elements. The dominant approach for doing so is to reduce the single multi-class problem into multiple binary classification problems. Nowadays, there are two common approaches for dealing with multi-class problems: (1) one of the labels and the rest (one-versus-all) or (2) between every pair of classes (one-versus-one). The one-versus-all strategy constructs one SVM binary classifier for each class by taking the samples in the considered class as the positive samples and all the other samples as the negative ones. The one-versus-one strategy constructs one SVM binary classifier for each pair of the classes, and the final classification is done in a max-wins voting way: every classifier

assigns the sample to one of the two classes, and the vote for the assigned class is then increased by one, and the sample is finally classified to the class with the most votes. Such strategy is adopted in C-SVC of the popular LibSVM implementation[Chang & Lin 2011].

2.4 Fusion strategies

In order to detect and annotate visual concepts from huge digital libraries, the needs for the combination (fusion) of several models are required. The idea of fusion is usually adopted in the problem of multimedia data analysis. For instance, an efficient fusion scheme must enhance concept indexing in multimedia documents by merging visual and textual modalities, color and texture modalities, or global and local features. Using a generic framework, usual approaches propose either to merge data on a concatenated vector before achieving classification, to perform several classification and then to merge confidence scores using a higher level classifier by the means of a stacking technique, or take advantage of merging modalities at kernel level. There are several different strategies for fusion:

- **Early fusion** The features which are extracted from the image and the text associated with the image are concatenated to build a single feature vector. This vector of concatenated features is used to compute the kernel function. Then feeded kernel function into a classifier for the final classification.

- **Late fusion** The models from each individual properties is first fed into a classifier to get its classification score, and the scores from all the models are then combined into the final score according to a certain criterion, such as mean, max, min, and weighted sum. Suppose $S_i, i = 1, \ldots, N$ represent the scores from N individual channels, the final score S_{fusion} can be obtained as fellows:

  - **Mean:** $S_{fusion} = \frac{1}{N} \sum_{i=1}^{N} S_i$
  - **Max:** $S_{fusion} = max(S_1, \ldots, S_N)$
  - **Min:** $S_{fusion} = min(S_1, \ldots, S_N)$

– **Weighted sum:** $S_{fusion} = \frac{1}{N}\sum_{i=1}^{N}(\omega_i * S_i)$, where ω_i is the weight for the $i-th$ properties.

while a late fusion at score level is reputed as a simple yet effective way to fuse features of very different nature for machine-learning problems, there are many improved lata fusion approaches proposed. [Tiberius Strat *et al.* 2012] propose Hierarchical late fusion, which automatically filters out irrelevant classifiers, then it groups highly-correlated ones in an iterative manner. [Liu *et al.* 2013] propose the selective weighted late fusion (SWLF), which selects and weights the scores from the best features according to the concept under hand to be classified.

- **Intermediate fusion** Different from both early and late fusion, Intermediate fusion method combines different features in the kernel level, and thus can be considered as a intermediate fusion strategy. Advantages of merging modalities at kernel level are numerous. First, it allows to choose the kernel functions according to the modalities. Second, kernel fusion also allows to model the data with more appropriate parameters[Ayache *et al.* 2007]. The Multiple Kernel Learning (MKL) method can also be interpreted as a kind of intermediate fusion technique. It has proven to be an extremely effective discriminative approach to classification as well as regression problems. Given multiple sources of information, one might calculate multiple basis kernels, one for each source. In such cases, the resultant kernel is often computed as a convex combination of the basis kernels,

$$K(x_i, x_j) = \sum_{m=1}^{M} d_m K_m(x_i, x_j), \quad \sum_{m=1}^{M} d_m = 1, \quad d_m \geq 0 \tag{2.36}$$

where x_i are the feature vectors, $K_m(x_i, x_j)$ is the $m-$th kernel and d_m are the weights given to each information source(kernel). Learning the classifier model parameters and the kernel combination weights in a single optimization problem is known as the Multiple Kernel Learning problem[Lanckriet *et al.* 2004]. For binary classification, given the learning set $\{x_i, y_i\}_{i=1}^{M}$, where x_i belongs

to some input data and y_i is the label of x_i, the decision function of canonical MKL is given as follows. The goal in SVM learning is to learn the globally optimal values of w and b from training data x_i, y_i. In addition,

$$f(x) = \sum_{i=1}^{N} \alpha_i^* y_i \sum_{m=1}^{M} d_m K_m(x_i, x) + b^* \tag{2.37}$$

where $\{\alpha_i^*\}_{i=1}^N$ and b^* are the coefficients of the classifier, corresponding to the lagrange multipliers and the bias in the canonical SVM problem. To solve the MKL problem efficiently, the SMO-MKL algorithm is used to optimise the l_p MKL dual[Vishwanathan *et al.* 2010].

The primal can therefore be formulated as

$$\begin{aligned} min \quad & \{\sum_k \frac{1}{d_m} w_k w_k^T + C \sum_i \xi_i\} \\ s.t. \quad & y_i \sum_k \phi_k(x_i) + y_i b \geq 1 - \xi_i \quad \forall i \\ & \xi_i \geq 0 \quad \forall i \\ & \sum_m d_m = 1, \quad d_m \geq 0 \quad \forall m \end{aligned} \tag{2.38}$$

where b is the bias, ξ_i is the slack afforded to each data point and C is the regularization parameter. The solution to the above MKL formulation is based on a gradient descent on the SVM objective value. An iterative method alternates between determining the SVM model parameters using a standard SVM solver and determining the kernel combination weights using a projected gradient descent method.

2.5 Datasets and Benchmarks

We introduce several standard datasets and popular benchmarks available for the Visual Concept Detection and Annotation task. They will be used to carry out experiments in the following chapters.

2.5.1 PASCAL VOC

The PASCAL Visual Object Classes (VOC) challenge.[2] is a benchmark in visual object category recognition and detection, providing the vision and machine learning communities with a standard dataset of images and annotation, and standard evaluation procedures. It consists of two components: (1) a publicly available dataset of images and annotations, together with standard evaluation procedures; and (2) an annual competition and workshop. Organized annually from 2005 to present, this challenge and its associated dataset has become accepted in computer vision and machine learning communities as a benchmark for visual object recognition and detection.

The goal of this challenge is to recognize objects from a number of visual object classes in realistic scenes (i.e. not pre-segmented objects). It is fundamentally a supervised learning problem in that a training set of labelled images is provided. The number of object classes considered was only 4 in the starting year of 2005, and then increased to 10 in 2006, and has further increased to 20 since 2007. The object classes that have been selected are:

- Person: person
- Animal: bird, cat, cow, dog, horse, sheep
- Vehicle: aeroplane, bicycle, boat, bus, car, motorbike, train
- Indoor: bottle, chair, dining table, potted plant, sofa, tv/monitor

There are two principal challenge tasks:

- Classification: For each of the twenty classes, predicting presence / absence of an example of that class in the test image.
- Detection: Predicting the bounding box and label of each object from the twenty target classes in the test image.

[2]http://pascallin.ecs.soton.ac.uk/challenges/VOC/

Besides the challenge organized in each year, the PASCAL VOC 2007 dataset has become a standard benchmark for evaluating object recognition and detection algorithms, because all the annotations were made available in 2007 by the organizers but since then they have not made the test annotations publicly available. The PASCAL VOC 2007 dataset contains nearly 10 000 images of 20 object classes, which contain different number of images, from hundreds to thousands. The dataset is divided into a predefined training set (2501 images), validation set (2510 images) and test set (4952 images). The mean average precision (mAP) across all the classes is used as the evaluation criterion. Average precision (AP) measures the area under the precision-recall curve for each class, and a good AP value requires both high recall and high precision values.

2.5.2 ImageCLEF

ImageCLEF.[3] launched in 2003 as part of the Cross Language Evaluation Forum (CLEF) with the goal of providing an evaluation forum for the cross-language annotation and retrieval of images. Motivated by the need to support multilingual users from a global community accessing the growing amount of visual information, ImageCLEF aims to support the advancement of the field of visual media analysis, indexing, classification and retrieval by developing the necessary infrastructure for the evaluation of visual information retrieval systems operating in both monolingual, cross-language and language-independent contexts. There are four main tasks in ImageCLEF:

- Robot vision Primary tabs
- Image Annotation
- Liver CT Annotation
- Domain adaptation

[3] http://www.imageclef.org/

date uploaded:	30 June 2008	
owner:	Shamima Desai	
username:	ShamsD	
license:	Attribution-NonCommercial-ShareAlike	
title:	Rich Chocolatey Chocolate Cake	
tags:	chocolate	
	cake	
	chocolateganachebuttercream	
	shamsd	
exif:	make = Nikon	aperture = f/5.6
	model = D40	max lens aperture = 5/1
	exposure = 0.033s	ISO Speed = 560
	exposure bias = 0/6EV	digital zoom ratio = 1/1
	focal length = 55mm	gain control = low gain up

Figure 2.16: An example photo from the MIRFLICKR collection and its associated user tags, user information, photo information, license information and EXIF metadata. Due to space considerations we only show part of the metadata.

2.5.2.1 Image Annotation

We participated in the ImageCLEF: Image Annotation challenge in 2010, and 2011. A brief introduction of our participation can be found in Charter 7. Automatic concept detection within images is a challenging and as of yet unsolved research problem. Impressive improvements have been achieved, although most of the proposed systems rely on training data that has been manually, and thus reliably labeled, an expensive and laborious endeavor that cannot easily scale. Recent image annotation benchmark campaigns have resorted to crowd source in order to label a large collection of images. However, when considering the detection of multiple concepts per image and an increasing list of concepts for annotation, even with crowd source the labeling task becomes too expensive. Thus, reducing the reliance on cleanly labeled data has become an necessity.

Dataset: The task uses a subset of the MIR Flickr 1 million image dataset for the annotation challenge. The MIR Flickr collection supplies all original tag data provided by the Flickr users. For most of the photos the EXIF data is included and may be used. One example is shown in figure 2.16.

Approach: The visual concept detection and annotation task is a multi-label classification challenge. It aims at the automatic annotation of a large number of consumer photos with multiple annotations.The task can be solved by following three different approaches:

- Automatic annotation with visual information only.

- Automatic annotation with Flickr user tags.
- Multi-modal approaches that consider visual information and/or Flickr user tags and/or EXIF information.

In this challenge, The Visual Concept Detection and Annotation(VCDA) is a multi-label classification challenge that offers a benchmark for testing novel visual concept detection, annotation and retrieval algorithms on a public collection containing photos gathered from the social sharing website Flickr. The aim is to analyze the images in terms of their visual and/or textual features in order to detect the presence of one or more semantic concepts. The detected concepts are then to be used for the purpose of automatically annotating the images or for retrieving the best matching images to a given concept-oriented query. The concepts are very diverse and range across categories such as people (e.g. teenager, female), scenery (e.g. lake, desert), weather (e.g. rainbow, fog) and even impressions (e.g. unpleasant, euphoric).

This task has a longstanding tradition at ImageCLEF. Since 2009 the task has been based upon various subsets of the MIRFLICKR collection, where every year the list of concepts to detect was updated in order to cover a wider selection of concept types and to make the task more challenging. The related PASCAL Visual Object Classes(VOC) challenge has as aim to accurately detect the bounding boxes and labels of objects in a set of images, whereas our focus is on both visual and textual information instead of visual information only and furthermore we offer a larger range of concepts to detect.

2.6 Conclusions

In this chapter, a review of multimodel approaches proposed in the literature for visual concept detection is presented. In particular, more attention in paid to build visual and textual features; fusion and classifier, because they have become the popular framework for the visual concept detection tasks nowadays. Typically, this kind of approach consists of three parts: (1)building visual model; (2) building textual

model; and (3) fusion and classification (machine learning) algorithms. The popular methods adopted for each of these parts are reviewed in detail respectively. Moreover, several fusion strategies for combining different features are also introduced.

We apply the multimdel approach based on visual and textual models for the visual concept detection task in this thesis, and we believe that the visual and textual descriptors fusion is key step. [Parikh & Zitnick 2010] have recently confirmed that visual descriptors play a key role. Meanwhile, [Guillaumin *et al.* 2010][Wang *et al.* 2009] have proven that text associated with images can build a surprisingly powerful descriptor and the visual concept detection tasks can benefit from text associated with images. Therefore, the following chapters of this thesis will focus on the visual and textual descriptors, and will propose several effective and efficient visual features and several novel textual features. Finally, in order to benefit from visual models and textual models, we apply the most popular fusion techniques, the Multiple Kernel Learning(MKL) appraoch.

Chapter 3

Visual Features

Contents

3.1 Encoding Local Binary Descriptors by Bag-of-Features with Hamming Distance

This work presents a novel method for encoding local binary descriptors for Visual Object Categorization (VOC). Nowadays, local binary descriptors, e.g. LBP and BRIEF, have become very popular in image matching tasks because of their fast computation and matching using binary bitstrings. However, the bottleneck of

applying them in the domain of VOC lies in the high dimensional histograms produced by encoding these binary bitstrings into decimal codes. To solve this problem, we propose to encode local binary bitstrings directly by the Bag-of-Features (BoF) model with Hamming distance. The advantages of this approach are two-fold: (1) It solves the high dimensionality issue of the traditional binary bitstring encoding methods, making local binary descriptors more feasible for the task of VOC, especially when more bits are considered; (2) It is computationally efficient because the Hamming distance, which is very suitable for comparing bitstrings, is based on bitwise XOR operations that can be fast computed on modern CPUs. The proposed method is validated by applying on LBP feature for the purpose of VOC. The experimental results on the PASCAL VOC 2007 benchmark show that our approach effectively improves the recognition accuracy compared to the traditional LBP feature.

3.1.1 Introduction

The advent of digital imaging sensors used in mobile phones and consumer-level cameras has produced a growing number of digital image collections. An appropriate categorization of image contents could help to have access to high-level information about objects contained in images and to efficiently manage such large collections. However, Visual Object Categorization (VOC) is one of the most challenging problems in computer vision community, mainly due to intra-class variations such as occlusion, clutter, viewpoint and lighting condition changes, which are typical in the real-world situations. Many approaches for VOC have been proposed in the literature, and the typical pipeline includes the following three steps[Chatfield *et al.* 2011]: (1) extraction of global or local image features (e.g. SIFT[Lowe 2004a], SURF[Bay *et al.* 2008], LBP[Ojala *et al.* 2002b], etc.); (2) encoding of the local features in an image descriptor (e.g. a histogram of the quantized local features), global features can be directly sent to classifiers; (3) classification of the image descriptor by certain machine learning algorithms (e.g. support vector machine, decision tree, etc.)[Chatfield *et al.* 2011]. For the first step, many local image descriptors have been proposed in the literature,

such as SIFT, Color SIFT[van de Sande *et al.* 2010], HOG[Dalal & Triggs 2005], DAISY[Zhu *et al.* 2011], and so on. For the second step, the purpose of the encoding is to transform large set of local descriptors into a compact global image representation. The Bag-of-Features (BoF) method[Csurka *et al.* 2004] is the most popular approach to do this. It is based on the idea of partitioning the local descriptor space into information points whose internal structure can be disregarded or parameterized linearly. More precisely, it consists of clustering local descriptors from each image and summarizing the distribution of these descriptors in the form of a signature composed of reprehensive cluster members and weights proportional to cluster sizes. The cluster centers are called visual words and the set of visual words is called a visual vocabulary. Many experimental results presented in the literature have clearly demonstrated that the BoF model is robust to background clutter and produces very good performances in the VOC tasks. The typical BoF method usually applies k-means algorithm for clustering and encodes local descriptors into global histograms by different encoding methods such as histogram encoding[Chatfield *et al.* 2011], kernel codebook encoding[Philbin *et al.* 2008, Gemert *et al.* 2008], fisher encoding[Perronnin *et al.* 2010], and so on. Finally, these encoded histograms are feeded into a classifier, e.g. SVM, to perform the classification.

Recently, local binary descriptors, e.g. LBP and BRIEF), are becoming increasingly popular in the computer vision domain. Compared to other popular local descriptors such as SIFT, HOG, SURF and so on, binary descriptors are very fast to compute and match, as well as possess advantages of memory and storage efficiency, because they are based directly on the binary bitstrings. They have exhibited good performances in image matching related tasks[Calonder *et al.* 2010]. However, the bottleneck of applying them in the domain of VOC lies in the high dimensional histograms produced by encoding these binary bitstrings into decimal codes. Let us take the LBP feature for example, which is introduced in chapter 2.

The final LBP feature consists of computing the LBP code for each pixel in an image and building a histogram based on these codes. Usually, considering bigger neighborhood (more neighboring pixels with bigger radius) could lead to better

performance because more local information is obtained. However, the drawback lies in the high dimensional histogram produced by the LBP codes. According to the definition, if the length of binary bitstring is p, the resulting histogram will be of 2^p dimension. The dimensionality growth is exponential when the number of neighboring pixels is increasing, and it is impractical to feed the histograms with such huge dimension into the classifier for classification.

In order to address this problem, instead of encoding the binary bitstrings into decimal codes, we propose to encode them directly by employing the BoF model with Hamming distance. The advantages are two-fold: (1) the dimensionality of the resulting histograms only depends on the size of the visual vocabulary, and is no longer related to the length of binary bitstrings, making local binary descriptors more feasible for the task of VOC, especially when more bits are considered; (2) It is computationally efficient because compared to other distance measurements such as Euclidean distance, the Hamming distance is more suitable for binary descriptors, and can be computed very efficiently via a bitwise XOR operation followed by a bit count. The proposed method will be validated in the experiments section by applying on LBP feature for the purpose of VOC.

The main contributions of this work are summarized as follows:

- Encoding local binary descriptors by the Bag-of-Features (BoF) model directly on binary bitstrings to address the high dimensionality issue and make them more feasible for the VOC tasks.

- Using Hamming distance together with k-means for visual vocabulary construction and histogram assignment for computational efficiency.

The remainder of the paper is organized as follows: In section 3.1.2, we present the proposed encoding method based on the BoF model with Hamming distance. In section 3.1.3, we describe our framework for the purpose of VOC. In section 3.1.4, we present the experimental results on the PASCAL VOC 2007 benchmark to validate the proposed approach. Finally in section 3.1.5, some conclusions are given.

3.1.2 Our Approach

Usually we get a local binary descriptor, which has a significant number of bits, from the neighborhood around one pixel. Instead of encoding the binary bitstrings into decimal codes, we would like to find a better way to make use of those bitstring descriptors. In this section, we propose to adopt the BoF model for encoding those bitstring descriptors. In the BoF model, two key steps include visual vocabulary construction and histogram assignment, where distance measurement plays an important role[Kumar & Annie 2012]. The chosen distance measurement determines how similar two elements are and how much time and computation resources are required. Here we propose to use Hamming distance[Vimal *et al.* 2008].

3.1.2.1 Hamming distance

In information theory, Hamming distance is named after Richard Hamming, its inventor, who introduced it in his fundamental paper on Hamming codes Error detecting and error correcting codes in 1950[Hamming 1950]. It is used in telecommunication to count the number of flipped bits in a fixed-length binary word as an estimate of error. For a fixed length bitstring, the Hamming distance is the number of positions at which the corresponding symbols are different. For binary bitstring the definition of the Hamming distance is as follows:

The Hamming distance (HD) $d(x, y)$ between two vectors $x, y \in F^n$ is the number of coefficients in which they differ, e.g.

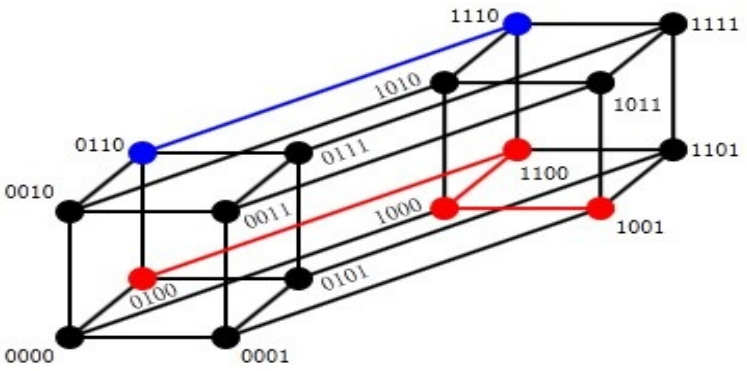


Figure 3.1: The example of the Hamming distance.

$$\text{in } F^4 \; d(0110, 1110) = 1 \quad \text{in } F^4 \; d(0100, 1001) = 3$$

The binary descriptors extracted from an image all have the same dimension and contain only 0 or 1. They are thus very efficient both to compute and to store in memory. Considering the computational efficiency, the Hamming distance in the BoF model is a better choice compared to the other measurements such as the Euclidean distance, because it can be computed extremely fast on modern CPUs that often provide a specific instruction to perform a XOR and a bit count operation.

3.1.2.2 Bags-of-Visual Words model with Hamming distance

Our motivation of using the BoF model with HD is to overcome rapidly increasing dimensions of histograms which are produced by encoding binary descriptor that multiply each binary bit with weights.

The Hamming k-clustering problem[Gaasieniec *et al.* 2000] is : Let Z_2^d be the set of all strings of length d over the alphabet $\{0, 1\}$. Given a binary descriptor set of observations $(x_1, x_2, ..., x_n)$, where each observation is a d-dimensional binary strings, and a positive integer $k < n$. The k-means clustering algorithm partition the n observations into k set $(S_1, S_1, \ldots, S_k)$, the cluster center of S_i is $\mu_i \in Z_2^d$, where $i \in \{1, \ldots, k\}$. Meanwhile, the cumulative approximation error is:

$$arg\ min \sum_{i=1}^{k} \sum_{x_j \in S_i} |x_j - \mu_i|_{HD} \tag{3.1}$$

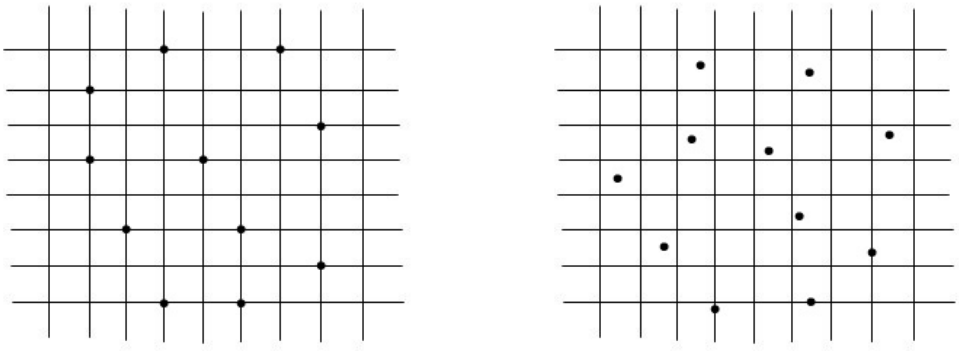
After we get the visual vocabulary $(\mu_1, \mu_2, ..., \mu_k)$, given a set of descriptors $(x_1, x_2, ..., x_n)$ extracted from an image, each descriptor x_i is assigned to the corresponding visual word according to:

$$arg\ min_{i=1...k} |x_j - \mu_i|_{HD} \tag{3.2}$$

Thus, each descriptor x_i is associated to a visual word μ_i, creating a histogram. The histogram encoded by the set of local descriptors is a non-negative vector F which is a k-dimensional vector. Finally each image can be represented by a non-negative vector F.

3.1.2.3 Comparison of BoF with Squared Euclidean distance

The performance of the BoF model depends on the k-means clustering. A cluster is a collection of data objects that are similar with objects within the same cluster and dissimilar to those in other clusters. Similarity between two objects is calculated using a single distance measurement[Vimal *et al.* 2008]. Choosing the right distance measurement for given data is very important[Shraddha Pandit 2011]. Moreover, there is a great deal of differences in computationally efficiency and computation resource economization. Fig. 3.2(a)(b) simply shows the Hamming space using the



(a) Cluster with HD

(b) Cluster with other distace

Figure 3.2: Illustration of k-means clustering with different distances. (a) k-means with HD, the centroids only appear in grid points. (b) k-means with Squared Euclidean distance, the centroids appear in the cell. Binary descriptors only appear on grid points, • is the centroids

2-dimension figure. The binary descriptor only appears in the grid points. The centroids calculated by k-means with HD appear in grid points. Compared with Squared Euclidean distance, the Hamming distance is better suitable for binary descriptor. If choosing Squared Euclidean distance, the centroids will appear in the cell. Fig. 3.2(a) shows k-means with HD. The centroids appear in grid points. Fig. 3.2(b) shows k-means with Squared Euclidean distance. The centroids appear in the cell.

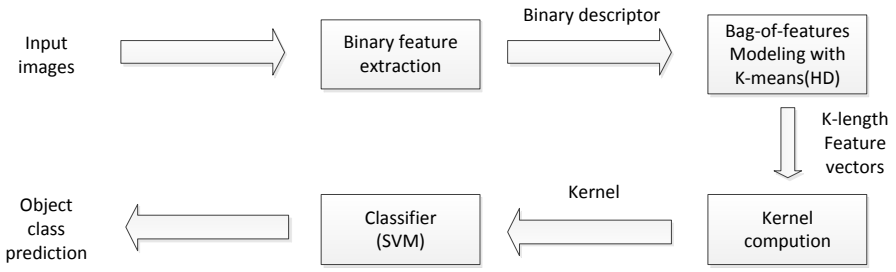


Figure 3.3: Flow chart of our system for VOC recognition

3.1.3 The Framework of VOC

Our framework for VOC is depicted in Fig. 3.3

3.1.3.1 Feature extraction

We choose the LBP descriptor to validate our approach. A brief introduction of LBP has been presented in the introduction section. The LBP descriptor is further extended to use the circular neighborhood with variant radius and variant number of neighboring pixels[1], as shown in Fig. 3.4.

Accordingly, the LBP bitstring at (x_c, y_c) is defined as follows:

$$\tau(g_c, g_p) = \begin{cases} 1 & \text{if} \quad g_c < g_p \\ 0 & \text{otherwise} \end{cases} , \tag{3.3}$$

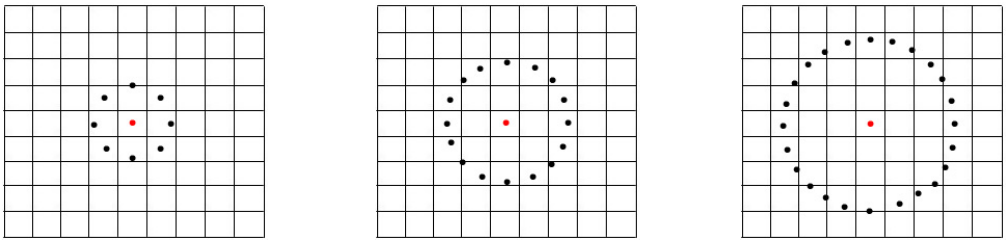
where g_p is the gray level value of the neighboring pixel, g_c is the value of the central pixel. According to the radius value, the number of neighbor pixels P is decided. We take the LBP descriptor to be the P-dimensional bitstring:

$$f_p(g_c) := \Sigma_{1 \leqq i \leqq p} 2^{\tau}(g_c, g_p) \quad . \tag{3.4}$$

In this paper, we employ LBP with the parameters $\{R = 1, P = 8\}$; $\{R = 2, P = 16\}$; $\{R = 3, P = 24\}$; $\{R = 4, P = 32\}$; $\{R = 5, P = 40\}$; $\{R = 6, P = 48\}$; and also extend LBP to the multi-scale form $\{R = 1, 2, 3 \quad P = 48\}$. This extension can get more local information around the central pixel. Finally, the LBP bitstring is

[1] http://www.cse.oulu.fi/CMV/Downloads/LBPMatlab

computed at every pixel location.



(a) R=1,P=8 (b) R=2,P=16 (c) R=3,P=24

Figure 3.4: Single-scale LBP operator

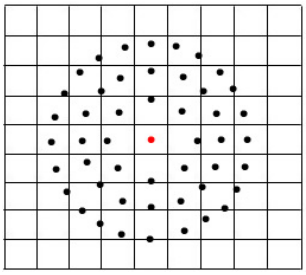


Figure 3.5: Multi-scale LBP

3.1.3.2 BoF model with Hamming distance

After feature extraction, each input image is represented by a set of LBP bitstrings. Compared with the decimal values of the descriptor vector, the values of the binary descriptor vector has only two values(0 or 1). We then adopt the BoF model with HD here to encode these bitstrings into global representation for each image, as presented in section 3.1.2.

3.1.3.3 Classification

Once all the BoF representations of the input images are obtained, they are then feeded into a certain classifier for classification. Here we apply the Support Vector Machine (SVM) for the final classification. The benefits of SVM for histogram-based classification have been clearly demonstrated in [Caputo *et al.* 2005].

In our experiments, the χ^2 distance is computed to measure the similarity between each pair of the feature vectors F and F' (n is the size of the feature vector):

$$dist_{\chi^2}(F, F') = \sum_{i=1}^{n} \frac{(F_i - F'_i)^2}{F_i + F'_i} \tag{3.5}$$

Then, the kernel function based on the χ^2 distance is used for SVM to train the classifier:

$$K_{\chi^2}(F, F') = e^{-\frac{1}{D} dist_{\chi^2}(F,F')} \tag{3.6}$$

where D is the parameter for normalizing the distances. Here D is set to the average distance of all the training data. Finally, for each test image, the output probabilities of SVM classifier are used to predict the object categories.

3.1.4 Experimental evaluation

We perform the VOC experiments on the standard PASCAL VOC 2007 benchmark. The dataset has 20 different object classes, such as sheep, train, boat, bus, sofa, table, etc. The dataset is pre-defined into 50% for training/validation and 50% for testing. In total there are 9,963 images, where 2501 are for training, 2510 are for validation and 4952 are for test.

All the images in the PASCAL VOC 2007 dataset come from the real world, thus yielding large variations in viewing and lighting conditions. Meanwhile, there also exist shape variations such as scaling and orientation of objects. All of these increase the difficulties of the VOC tasks on this dataset. For evaluation we use mean average precision (mAP)[Yue *et al.* 2007]. i.e., for each test categoey we obtain a precision/recall curve, and then compute its average precision based on the area under this curve. Finally the mean value over all the categories is computed.

3.1.4.1 Experimental setup

In order to validate the proposed approach, we compared the performance of mAP obtained by the original LBP feature and the feature using the BoF model with HD.

(a) Aeroplane

(b) Dog

(c) Horse

(d) Bicycle

(e) Car

(f) Tv/monitor

(g) Bird

(h) Cat

Figure 3.6: Example images of the PASCAL VOC 2007 benchmark.

Moreover, we also compared the time consumption of the BoF model with different distance measurements. The fact that the experiments with different distances consist of similar steps allows us to make the time consumption comparison in k-means and the assignment step.[2]

As we described in section 3.1.3, for feature extraction of LBP, there are 2 main parameters need to be decided: the size of neighborhood, defined as radius R ; and the number of neighboring pixels, defined as P, that are taken into account on the circle of radius R. We use the original LBP implementation available online.[3] For our approach, the length of the binary descriptor is equal to P. In the BoF modeling step, the factor that must be decided during the experiments is the size of the visual vocabulary, defined as C. The experimental results in the literature have clearly demonstrated that larger vocabulary leads to better performances[Chatfield *et al.* 2011]. But too big size of the vocabulary will also make the resulting histograms too sparse. To find a good size of visual vocabulary, we have made a series of experiments and chosen the optimization size for vocabulary as follows: for $\{R = 1, P = 8\}, C = 220$; for $\{R = 2, P = 16\}, C = 1300$; for $\{R = 3, P = 24\}, C = 1400$. The LibSVM implementation[Chang & Lin 2011] of the SVM is used to train the classifier.

3.1.4.2 Comparison of our approach with original LBP

The mAP results of our approach on the PASCAL VOC 2007 benchmark are shown in Fig 3.7. It can be seen that the performances of encoding the bitstrings using our approach are better than the original LBP which uses the multiplying weights for encoding step. More specifically, we can observe that: (1) in the case of $\{R = 1, P = 8\}$, our approach gets a performance with mAP of 28.40%, which is comparable and somewhat better than the performance of mAP 28.30% of the original LBP[Zhu *et al.* 2011]; (2) in the case of $\{R = 2, P = 16\}$, our approach gets the best performance with mAP of 33.09% , which is also better than the original LBP approach; (3) in the case of $\{R = 3, P = 24\}$, the results of our approach are

[2]We use the MATLAB implementations available for k-means and C++ for assignment.
[3]http://www.cse.oulu.fi/CMV/Downloads/LBPMatlab

still better than the original LBP approach.

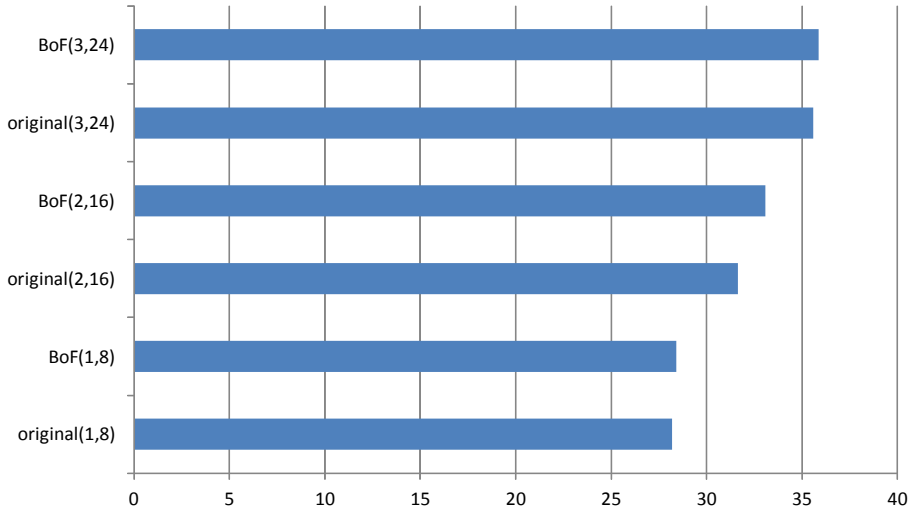


Figure 3.7: Comparison of different LBP scale and the number of points in terms of original(R,P) and BoF(R,P) with HD approach and classification accuracy on PASCAL 2007 (P, R: P neighboring pixels equally located on a circle of radius R)

Because of the bottleneck that lies in the high dimensional histogram produced by encoding LBP by multiplying the bits with weights, it is almost impossible to encode LBP where the number of neighboring pixels is above 32. In Table 3.1, the results proved that our approach can solve this problem. It also can be observed that the large radius LBP can still capture discriminative information. For example, when {R=4,P=32}, codebook size 1400; {R=5,P=40}, codebook size 1200; {R=6,P=48}, codebook size 1200, the performances of mAP are all above 32%.

Table 3.1: The performance of classification accuracy on PASCAL 2007 using BoF(R,P) with HD approach (P, R: P neighboring pixels equally located on a circle of radius R)

BoF{R,P}	mAP(%)
BoF{R=4,P=32}	32.52
BoF{R=5,P=40}	33.24
BoF{R=6,P=48}	32.78

3.1.4.3 Comparison between multi-scale binary descriptor and multi-scale fusion

We also evaluated the LBP bitstrings after multi-scale fusion. We directly combine multi-scale LBP by concatenating the bitstrings from different scales. In our experiments, we extracted {R=1,2,3 P=8+16+24} for each pixel. The length of binary feature after fusion is 48. Meanwhile we compare with the traditional LBP fusion approach which fuses $\{R = 1, P = 8\}$ $\{R = 2, P = 16\}$ $\{R = 3, P = 24\}$ in histogram level. The comparison results are shown in Table 3.2. It can be seen that the multi-scale LBP using our approach gets the performance of mAP 35.17%. Compared with the traditional LBP fusion approach, our approach obtains an interesting performance improvement (nearly 2%).

Table 3.2: Comparison between multi-scale fusion(MSF) $\{R = 1, 2, 3\}$ in histogram level and multi-scale binary(MSB) $\{R = 1, 2, 3 \quad P = 8+16+24\}$ on PASCAL VOC 2007(P, R: P neighboring pixels equally located on a circle of radius R)

Multi-scale{R,P}	mAP(%)
MSF{R=1,2,3}	32.49
MSB{R=1,2,3 P=8+16+24}	35.17

3.1.4.4 Comparison with other Texture Descriptors

As one kind of texture feature, LBP(BoF) are compared with three widely-used popular texture descriptors, including Gabor filter, Texture Auto Correlation (TAC), and Grey Level Co-occurrence Matrix (GLCM). We set 5 scales and 8 orientations for Gabor filter. For TAC, the range of x and y directions is [0, 8] with interval of 2. For GLCM, 4 directions (horizontal, vertical and diagonal) with 1 offset between two pixels are considered.

From the results shown in Fig 3.13, it can be seen that the original LBP already outperforms other popular texture descriptors, proving that LBP is one of the best texture features available today. Our LBP(BoF) approach further improve the performances to almost double of the other texture descriptors, demonstrating that the strong power of our approach.

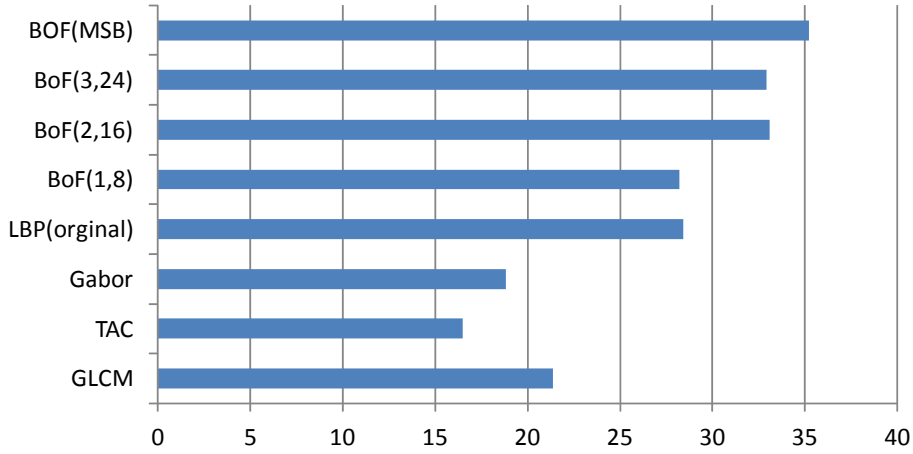


Figure 3.8: Comparison between BoF(R,P) with Hamming distance with other texture descriptors classification accuracy on PASCAL 2007 (P, R: P neighboring pixels equally located on a circle of radius R)

3.1.4.5 Comparison of the computational cost

A good approach should be both computationally efficient and computation resource economized. Compared to the original LBP encoding approach, our approach costs more time. In order to reduce the time consumption, we employ the HD for clustering and assignment. Here the comparison of the computational cost between the HD and the Euclidean distance is shown in Table 3.3. The comparisons are conducted on an Intel(R) Core(TM) i7 CPU 940 @ 2.93GHz with 9GB RAM. It can be seen that the time consumption of k-means with Squared Euclidean distance is almost 200 times than k-means with HD. The assignment of each image with HD is faster than the assignment with Squared Euclidean distance. The performances of mAP between these two distances are very close, as shown in Table 3.4.

Table 3.3: Comparison of Computation times for Hamming distance and Squared Euclidean distance in k-means step and assignment steps on PASCAL 2007 ($P = 8, R = 1$: P neighboring pixels equally located on a circle of radius R)

Time (second)	k-means	assignment(each image)
Hamming distance	98.93	1.74
Squared Euclidean distance	19046.43	2.76

Table 3.4: Comparison between BoF with Hamming distance and BoF with Squared Euclidean distance classification accuracy on PASCAL 2007 ($P = 8, R = 1$: P neighboring pixels equally located on a circle of radius R)

Distance	mAP(%)
Hamming distance	28.20
Squared Euclidean distance	29.13

3.1.5 Conclusions

In this work, we introduced a novel approach to use local binary descriptors for the task of VOC. The main contributions are to propose a new encoding method to address the high dimensionality issue of the traditional binary bitstring encoding, and to adopt Hamming distance with the BoF model for visual vocabulary construction and histogram assignment. HD is suitable for computer instruction because it performs an XOR operation. In contrast to other distances, HD spends less time and needs less computer resource.

The proposed approach was validated by applying on the LBP feature on the PASCAL VOC 2007 dataset. Compared with the original LBP, it exhibited better recognition accuracy. Meanwhile we extended the LBP to multi-scale form by directly concatenating binary bitstrings, and also obtained better performance than the traditional multi-scale fusion in histogram level. The time consumption is very reasonable. Compared with encoding LBP, the binary LBP also has the same property with original LBP.

Future work could consider to use other local binary descriptors (e.g. BRIEF) in our framework for the task of VOC as well as texture classification. Moreover, the proposed approach can be extended to different color spaces (e.g. HSV and OPPONENT) to improve the performance.

3.2 Sampled Multi-scale Color Local Binary Patterns

In this part, we propose a novel representation, called sampled multi-scale color Local Binary Pattern (SMC-LBP), and apply it to Visual Object Classes (VOC) Recognition. The Local Binary Pattern (LBP) has been proven to be effective for

image representation, but it is too local to be robust. Meanwhile such a design cannot fully exploit the discriminative capacity of the features available and deal with various changes in lighting and viewing conditions in real-world scenes. In order to address these problems, we propose SMC-LBP, which randomly samples the neighboring pixels across different scale circles, instead of pixels from individual circular in the original LBP scheme. The proposed descriptor presents several advantages. The experimental results on the PASCAL VOC 2007 image benchmark show significant accuracy improvement by the proposed descriptor compared with both the original LBP and other popular texture descriptors.

3.2.1 Introduction

Texture, color and local gradients features play a major role in content-based image categorization task. Identifying patches with texture features is at the heart of many computer vision algorithms. It is widely applied in object category recognition and image retrieval application[Ozuysal *et al.* 2010]. Identifying patches is difficult because of drastic surface appearance which depends on how the image texture information is captured. To address this problem, many texture descriptors have been proposed in the literature, such as Grey Co-occurrence Matrix(GLCM)[Tuceryan & Jain 1998], Texture Auto Correlation(TAC)[Tuceryan & Jain 1998], Gabor filter[Zhang *et al.* 2000], Brief[Calonder *et al.* 2010] and LBP[Ojala *et al.* 2002a].

Among all these texture features, LBP is one of the most popular texture descriptors. It was introduced in Chapter 2. The LBP descriptor is further extended to multi-scale using a circular neighborhood with variant radius and variant number of neighboring pixels. Fig 3.4 illustrates circularly symmetric neighbor sets for various (P, R). The gray values of neighbors are estimated by interpolation.

Because of its descriptive power for analyzing both micro and macro texture structures, and computational simplicity, LBP has been widely applied for texture classification[Ojala *et al.* 2002a] and object recognization[Zhu *et al.* 2010], and has demonstrated excellent results and robustness against global illumination changes. It has also been used successfully for texture segmentation[Blas *et al.* 2008], recog-

nition of facial identity[Guo *et al.* 2010] and expression[Shan *et al.* 2009].

However, the original LBP descriptor also has several drawbacks in its application. It covers a small spatial support area, hence the bit-wise comparisons are made through single circular pixel values with the central pixel value. This means that the LBP codes are easily affected by noise[Liao *et al.* 2007]. Moreover, features calculated in a single circular neighborhood cannot capture larger scale structure (macrostructure) that may be dominant features. Meanwhile, the original LBP descriptor ignores all color information (its calculation is based on gray image), while color plays an important role for distinction between objects, especially in natural scenes[Zhu *et al.* 2010]. There can be various changes in lighting and viewing conditions in real-world scenes, leading to large variations of objects in surface illumination, scale, etc., which make the original LBP performance is not very good in VOC recognition tasks. In order to address these drawbacks, many improve method of LBP descriptors have been proposed, such as Multi-scale Block LBP[Liao *et al.* 2007], Hierarchical Multi-scale LBP[Guo *et al.* 2010], Multi-scale Color LBPs[Zhu *et al.* 2010] and so on.

Traditionally, in order to capture larger scale structure (macrostructure), the histogram fusion and extending radius approaches have been proposed. Firstly, LBP features of different scale are extracted, and then the histograms are concatenated into a long feature. Vector joint distribution could contain more information, but it suffers from huge feature dimension. Meanwhile a single histogram can not represent a complete image content. Usually, considering bigger neighborhood (more neighboring pixels with bigger radius) could lead to better performance because more local information is obtained. However, the drawback lies in the high dimensional histogram produced by the LBP codes. According to the definition, if the length of binary bitstring is p, the resulting histogram will be of 2^p dimension. The dimensionality growth is exponential when the number of neighboring pixels is increasing, and it is impractical to feed the classifiers with such huge dimension histograms for classification. Although many approaches reduced the dimension(e.g. ri, u2[Ojala *et al.* 2002a]) were proposed, the drawback are still not solved completely.

In this work, we propose a novel representation, called Sample Multi-scale Color

Local Binary Pattern (SMC-LBP), to overcome the mentioned limitations of LBP and extend the LBP feature to patch. To validate the proposed feature, we apply it to VOC Recognition problem. In SMC-LBP, the computation is done based on randomly sampling the neighboring pixels from multi-scale circles. Furthermore, in order to enhance photometric invariance property and discriminative power, the proposed descriptor is computed in different color spaces. To summarize, the SMC-LBP descriptor presents several advantages:

- It encodes not only single scales but also multiple scale of image patterns, extends the LBP to a patch, and hence provides a more complete image representation than the original LBP descriptor.

- It corporates with color information, therefore its photometric invariance property and discriminative power are enhanced.

In section 3.2.2, we introduce Sample Multi-scale Binary Pattern in detail. Section 3.2.3 presents Sample Multi-scale Color Local Binary Pattern. The Framework of the experiment is introduced in section 3.2.4. The experimental results are shown in section 3.2.5. Finally some conclusions and future work are given in section 3.2.6.

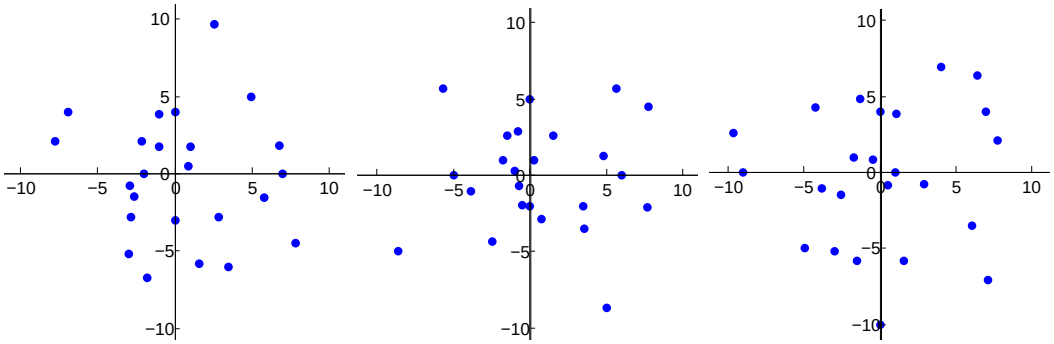
3.2.2 Sample Multi-scale Local binary pattern

3.2.2.1 SM-LBP Approach

Our approach is inspired by earlier work[Ozuysal *et al.* 2010] that demonstrates that image patches could be effectively classified on the basis of a relatively small number of pairwise intensity comparisons[Calonder *et al.* 2010]. Here we randomly sample across different scale circles, as shown in Fig 3.9 and is further extended to use the circular neighborhood with variant radius and variant number of neighboring pixels.

More specifically, the SM-LBP descriptor at pixel location g_c (x_c, y_c) is defined as follows:

$$SM - LBP_N := \Sigma_{n=1}^{N} \tau(g_c, g_n) 2^n \tag{3.7}$$



(a) Uniform$(-10, 10)$ (b) Gaussian$(0, 100)$ (c) Gaussian$(4, 100)$

Figure 3.9: Different approaches to choosing the g_n location. All the radius R are selected by randomly sampling from different circulares.

$$\tau(g_c, g_n) = \begin{cases} 1 & \text{if} \quad g_c < g_n \\ 0 & \text{otherwise} \end{cases} \tag{3.8}$$

where $g_n \sim (g_1, g_2, \cdots, g_N)$. g_n is the pixel gray value of the multi-scale circular neighborhood. N is the number of neighbor pixels which we randomly choose from different scale circles. How to generate the g_n is introduced in the next section.

Compared to the original LBP, the SM-LBP replaces comparisons between the central pixel and single circular pixels with comparisons between the central pixel and the pixels which are randomly chosen from multi-scale circles. In this way, the neighboring pixels randomly chosen could come from the different scales, this means that our new descriptor can capture more information from larger region. In this paper, the following experiments consider $N = 8$, 16, 24; g_n belongs to $\{R_1, R_2, R_3, R_4, R_5\}$, R_i is $i - th$ radius.

3.2.2.2 Sample Arrangement of SM-LBP

There are many options for generating the radius R_n from different distributions. We experiment with three sampling approaches. In the following, we assume that the origin of the patch coordinate system is located at the patch center. The patch size S is $max(R_i)$. The center point g_c (x_c, y_c) is located at the patch center. g_n (x_n, y_n) are given by $(-R_n \sin(2\pi n/N), R_n \cos(2\pi n/N))$, R_n can be described as

follows.

- $R_n \sim$ i.i.d. Uniform$(-R, R)$: The g_n locations are evenly distributed over the patch, as is shown in Fig.3.9(a).

- $R_n \sim$ i.i.d. Gaussian$(0, S^2)$,the radius R_n is sampled from a Gaussian distribution with mean parameter 0 and standard deviation parameter S^2 centered around the origin g_c. This forces the g_n to be more local. g_n locations outside the patch are clamped to the edge of the patch, as is shown in Fig.3.9(b).

- $R_n \sim$ i.i.d. Gaussian(R_i, S^2) , the radius R_n is sampled from a Gaussian distribution with mean parameter $R_i \neq 0$ and standard deviation parameter S^2 centered around the origin g_c. g_n locations outside the patch are clamped to the edge of the patch, as is shown in Fig.3.9(c).

3.2.3 Sample Multi-scale Color Local Binary Pattern

3.2.3.1 Model Analysis for Illumination Changes

The VOC task is important to access visual information on the level of objects and scene types[van de Sande *et al.* 2010]. In order to enhance the descriptor's illumination invariance and discriminative power, we further proposed color SM-LBP, called SMC-LBP. The diagonal model eq. (3.9) and the diagonal-off model eq. (3.10) can be used to model changes in the illumination[van de Sande *et al.* 2010].

$$\begin{pmatrix} R^c \\ G^c \\ B^c \end{pmatrix} = \begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix} \begin{pmatrix} R^u \\ G^u \\ B^u \end{pmatrix} \tag{3.9}$$

$$\begin{pmatrix} R^c \\ G^c \\ B^c \end{pmatrix} = \begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix} \begin{pmatrix} R^u \\ G^u \\ B^u \end{pmatrix} + \begin{pmatrix} O_1 \\ O_2 \\ O_3 \end{pmatrix} \tag{3.10}$$

where u is a light source, and c is the canonical illumination. The eq. (3.10) presents mAPs colors that are taken under an unknown light source to their corresponding

colors under the canonical illumination[Ozuysal *et al.* 2010]. In order to deal with a wider range of imaging conditions, Finlayson et al extend the diagonal model to the diagonal-off model with an offset $(O_1, O_2, O_3)^T$[Finlayson *et al.* 2005].

Based on above two models, illumination change can be defined. If a constant factor is in all channels ($a = b = c$) In eq. (3.9), it presents the light intensity change; If Image values change by an equal offset in all channels ($a = b = c = 1, O_1 = O_2 = O_3$) in eq. (3.10), it presents light intensity shift. If ($a = b = c, O_1 = O_2 = O_3$) in eq. (3.10), it means light intensity change and shift. Light color change depends on all channels independently ($a \neq b \neq c$), as eq. (3.9) and light color change depends on all channels independently with arbitrary offsets ($a \neq b \neq c, O_1 \neq O_2 \neq O_3$), as eq. (3.10).

3.2.3.2 SMC-LBP Descriptors

In order to enhance SM-LBP's photometric invariance property and discriminative power, three color SMC-LBP descriptors are proposed. The main idea is to compute the SMC-LBP descriptor independently over all the channels of certain color spaces.

RGB-SM-LBP This descriptor is obtained by computing LBP over all three channels of the RGB color space independently, and then concatenating the results together. It is invariant to monotonic light intensity change due to the property of the original LBP, and has no additional invariance properties.

Opponent-SM-LBP This descriptor is obtained by computing LBP over all three channels of the opponent color space:

$$\begin{pmatrix} O_1 \\ O_2 \\ O_3 \end{pmatrix} = \begin{pmatrix} (R-G)/\sqrt{2} \\ (R+G-2B)/\sqrt{6} \\ (R+G+B)/\sqrt{3} \end{pmatrix} \tag{3.11}$$

Due to the subtraction, O_1 and O_2 channels are invariant to light intensity shift. O_3 channel represents the intensity information, and has no invariance properties.

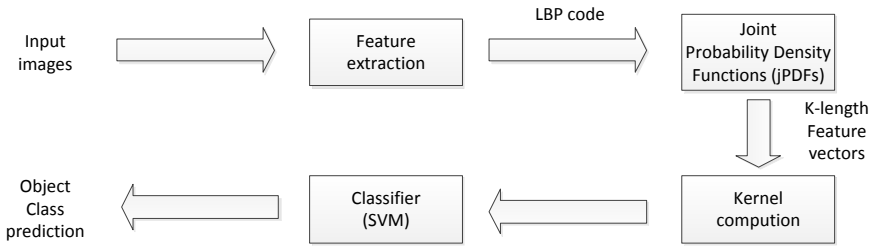


Figure 3.10: Flow chart of our system for VOC.

Hue-SM-LBP This descriptor is obtained by computing LBP for the Hue channel of the HSV color space:

$$Hue = arctan(\frac{O_1}{O_2}) = arctan(\frac{\sqrt{3}(R-G)}{R+G-2B}) \tag{3.12}$$

Due to the subtraction and the division, Hue channel is scale-invariant and shift-invariant, therefore this descriptor is invariant to light intensity change and shift.

3.2.4 The Framework of VOC

Our framework for VOC is depicted in Fig. 3.10

3.2.4.1 Feature Extraction

The SM-LBP descriptors extracted from input images at every pixel location as their features. With the radius R_n which is sampled from a Gaussian or Uniform distribution, the neighboring pixels g_n (x_n, y_n) are generated. By this way, the LBP descriptor is extended to use the multi-circular neighborhood with variant radius and variant number of neighboring pixels. It is more suitable for VOC task. Moreover, in order to increase photometric invariance property and discriminative power of the SM-LBP descriptors, The SMC-LBPs are proposed and used in this system.

3.2.4.2 Classification

Once all the joint probability density functions (jPDFs)[Lategahn *et al.* 2010] representations of the input images are obtained, they are then feed into certain classifier

(a) Aeroplane (b) Dog (c) Horse

(d) Bicycle (e) Car (f) TV/monitor

Figure 3.11: Example images of the PASCAL VOC 2007 benchmark.

for classification. Here we apply the Support Vector Machine (SVM) for the final classification. The benefits of SVM for histogram-based classification have been clearly demonstrated in [Caputo *et al.* 2005].

In our experiments, the χ^2 distance is computed to measure the similarity between each pair of the feature vectors F and F' (n is the size of the feature vector):

$$dist_{\chi^2}(F, F') = \sum_{i=1}^{n} \frac{(F_i - F'_i)^2}{F_i + F'_i} \tag{3.13}$$

Then, the kernel function based on the χ^2 distance is used for SVM to train the classifier:

$$K_{\chi^2}(F, F') = e^{-\frac{1}{D} dist_{\chi^2}(F,F')} \tag{3.14}$$

where D is the parameter for normalizing the distances. Here D is set to the average distance of all the training data. Finally, for each test image, the output probabilities of SVM classifier are used to predict the object categories.

3.2.5 Experiment

We perform the VOC experiments on the standard PASCAL VOC 2007 benchmark. The dataset has 20 different object classes, such as sheep, train, boat, bus, sofa, table, etc. Some example images are shown in Fig. 3.11. The dataset is pre-defined into 50% for training/validation and 50% for testing. In total there are 9,963 images, where 2501 are for training, 2510 are for validation and 4952 are for test.

For evaluation we use mean average precision (mAP)[Yue *et al.* 2007]. We train the classifier on the training set, then tune the parameters on the validation set, and obtain the classification results on the test set. The mAP is computed based on the proportion of the area under this curve.

3.2.5.1 Experimental Results

In order to evaluate the performance of our descriptors, we compare SM-LBP and SMC-LBP descriptors with the other texture features. Meanwhile we also compare these descriptors with the SIFT[Lowe 2004a] which is one of the most powerful image descriptors.

3.2.5.2 Comparison with the Original LBP

The proposed SM-LBP descriptors are compared with the original LBP. In our experiment, we set to $N = 8, 16, 24$, and g_n are generated by the Gaussian distribution and the Uniform distribution. The final mAP value is obtained by the mean of 20 experimental results. Table 3.5 shows the comparison of proposed SM-LBP descriptors and the original LBP on PASCAL 2007. It can be seen that the SM-LBP gets the better performance of mAP. Compared with the original LBP, the SM-LBP obtains a better performance improvement (nearly 2%). Fig. 3.12 shows comparison of the proposed SMC-LBP descriptors and original color LBP. It shows that the SMC-LBP all further outperform the original color LBP, with the improvements from 2% to 5.8%.

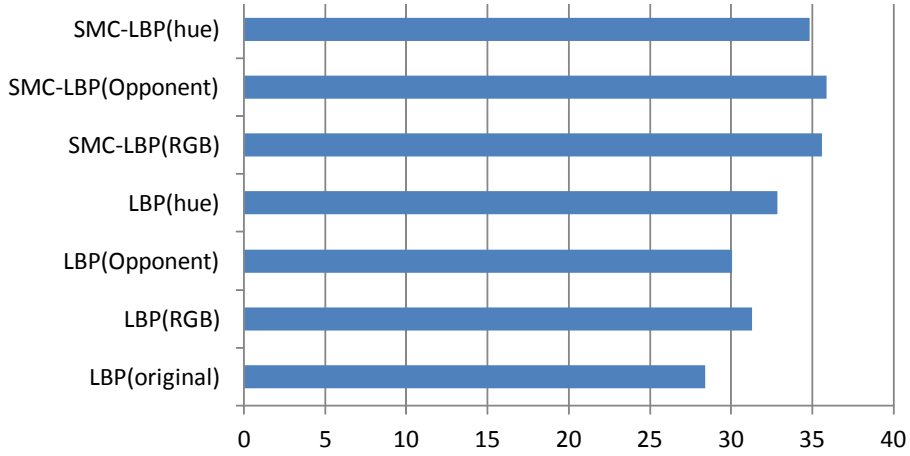


Figure 3.12: Comparison of the proposed SMC-LBP descriptors and original color LBP(For original color LBPs, N=24, R=3; For the SMC-LBPs, N=24, the distribution is Gaussian(0, 25)).

Table 3.5: Comparison of proposed SM-LBP descriptors and the original LBP on PASCAL 2007(original LBP: N=8, the circle of radius R=1; N=16, the circle of radius R=2; N=24, the circle of radius R=3. U,G: U is the Uniform distribution; G is the Gaussian distribution;).

mAP(%)	N=8	N=16	N=24
LBP(original)	28.40	31.64	29.78
SM-LBP(U(-5,5))	30.40	33.83	33.02
SM-LBP(G(0, 25))	30.61	33.42	33.20
SM-LBP(G(2, 25))	29.98	33.34	33.32

3.2.5.3 Comparison with other Texture Descriptors

As one kind of texture feature, SM-LBP and SMC-LBP are compared with three widely-used popular texture descriptors, including Gabor filter, Texture Auto Correlation (TAC), and Grey Level Co-occurrence Matrix (GLCM). We set 5 scales and 8 orientations for Gabor filter. For TAC, the rang of x and y directions is [0,8] with interval of 2. For GLCM, 4 directions (horizontal, vertical and diagonal) with 1 offset between two pixels are considered.

From the results shown in Fig 3.13, it can be seen that the original LBP already outperforms other popular texture descriptors, proving that LBP is one of the best texture features available today. Our new descriptors further improve the performances to almost double of the other texture descriptors, demonstrating that the strong power of the proposed descriptors benefit from the properties of illumination-invariant and scale-invariant.

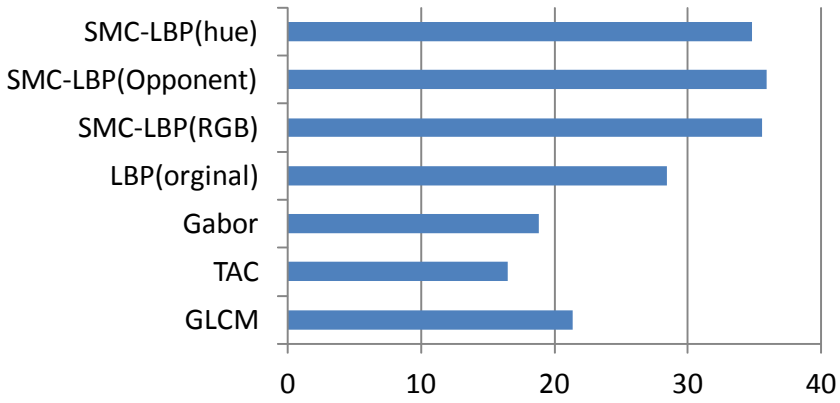


Figure 3.13: Comparison of the proposed SMC-LBP descriptors and other texture descriptors(SMC-LBPs, N=24, the distribution chosen Gaussian(0, 25)).

3.2.5.4 Comparison with SIFT Descriptor

Nowadays SIFT, a kind of local gradient descriptors is one of the most powerful image descriptors in the literature. Comparison of the proposed SMC-LBP and the SIFT, shows that the performance of our texture SMC-LBP descriptor is close to SIFT.

Table 3.6: Comparison of the proposed SMC-LBP and the SIFT(SMC-LBPs, N=24, the distribution chosen Gaussian(0, 25)).

	mAP(%)
LBP(original)	28.40
SMC-LBP(hue)	34.82
SMC-LBP(Opponent)	35.87
SMC-LBP(RGB)	35.59
SIFT	38.00

3.2.6 Conclusions

In this chapter, we propose a novel SM-LBP descriptor which can obtain multi-scale patterns and provide a patch texture representation. Moveover, in order to deal with the deficiency of color information and sensitivity to non-monotonic lighting condition changes, SMC-LBP descriptor is proposed. The main contributions are that SM-LBP and SMC-LBP not only have more discriminative power by obtaining more local information, but also possess invariance properties to different lighting condition changes. In addition, they keep the advantage of computational simplicity from the original LBP descriptor. The proposed descriptors are validated by applying on on the PASCAL VOC 2007 image benchmark. Compared with the original LBP and other texture descriptors, the experimental results exhibit better recognition accuracy.

Chapter 4

Textual Features

Contents

4.1 Introduction

With the advent of the digital camera and the popularity of internet photo sharing sites, more and more images are shared on internet. These images are usually annotated by users with tags or keywords. How can we use these annotations to help us detect and annotate new images?

The main idea of this chapter is to use tags associated with images to build textual features to automatically detect and annotate images. Usually typical tags associated with images include two following kinds of styles: (1) Text from web pages; (2) Manual annotation. The Visual Concept Detection and Annotation(VCDA) task is a multi-label classification challenge. It aims at deciding

Tags: paris iledefrance france
Label: Indoor Person Adult calm

Tags: colours cores comercial londrina paraná diogo figueira
Lables: Macro Outdoor Citylife Day Clouds Sky

Tags: puppy dog mini hands
Label: Plants Animals cute dog

Tags: soldier aircraft apache helicopter
Label: Desert Outdoor Portrait

Tags: outdoors newlife nature native mothers cute baby
Label: Outdoor Plants Sunny Animals cute

Tags: downtown millionaire Houston
Label: Outdoor Vehicle boring

Figure 4.1: Example images form ImageCLEF 2011data sets with their associated tags and class labels.

whether a large number of images which come from consumers belong to a certain concept. These images usually sense (e.g. Vehicle, Animals, Plants, etc), events (e.g. travel, work, etc.), or even sentiments (melancholic, cute, happy, funny, etc.). Due to large intra-class variations and inter-class similarities, clutter, occlusion and pose change[Guillaumin *et al.* 2010], this work is extremely challenging in computer vision domain.

The state-of-the-art methods in computer vision community have so far focused more on visual content descriptor and less on textual descriptor. This may be because tags associated with images are tended to be noisy in the sense so that they are not directly related to the pixel information but more to the semantic of the image content. Thus there is still much information in tags, as shown in Figure 4.1. This kind of information is hard to describe by visual descriptors. Usually the term frequencies model is used to represent the tags as bag-of-words(BoW), where each component of the vector is word count or term frequency. The BoW approach achieves good performance on the VCDA task. However this approach has two main drawbacks:

- The BoW is sensitive to the changes in vocabulary that occur when training data can not be reasonably expected to be representative of all the potential testing data.
- The BoW only considers the word frequency information, thus disregards tags semantic information.
- The performance of the BoW seriously depends on dictionary construction.

Semantic distance We relay on the $WordNet$ to measure the distance between two words. $WordNet$ structure[Fellbaum 1998] can be seen as a semantic network where each node represents a concept of the real world. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. These synsets are connected by arcs that describe relations between concepts. The semantic similarity between w_1 and w_2 is defined by:

$$SIM(w_1, w_2) = \begin{cases} sim(s_1, s_2) & \text{if a } CS \text{ exists for } s_1 \text{ and } s_2 \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

$$sim(s_1, s_2) = \frac{\min\{lcs(s_1), lcs(s_2)\}}{depth(CS) + \min\{lcs(s_1), lcs(s_2)\}} \tag{4.2}$$

where s is a synset and $w_i \in s_i$. $lcs(s)$ denotes the distance form s to the common subsume (CS) (most specific ancestor node) of the two synsets s_1 and s_2 in a $WordNet$ taxonomy. $depth(CS)$ is the length of the path from CS to the taxonomy $Root$[1].

Semantic Textual Feature using a Dictionary In order to solve these problems, we propose the semantic BoW model that uses the textual semantic information to build the semantic features. How to estimate the semantic similarity between words is one of the longest-established tasks in nature language processing and many approaches have been developed. In our approach the semantic distance between tags is measured based on their relative and absolute position in a graph such as the $WordNet$ hierarchy[Fellbaum 1998]. The $WordNet$ lexical hierarchy is used to build text semantic feature. Thus we expect to have a more precise access to the high level text semantic information contained in tags than what the text frequencies analysis gives.

The main contributions of this work are summarized as follows:

- We build textual descriptors by the semantic BoW model, in order to capture the semantic information between tags which is hardly described by the term frequencies model.

- We use $WordNet$-based semantic distance for dictionary construction and histogram assignment, in order to reduce the size of the tags representation.

Semantic Textual Feature without Dictionary In order to solve that the BoW only considers the word frequency information, disregards tags semantic information, Ningning Liu et al[Liu *et al.* 2011a] propose that building textual feature

[1]http://rednoise.org/rita/wordnet/documentation/riwordnet_method_getdistance.htm

based on *WordNet* distance for VCDA task and demonstrate that it especially improves performance of VCDA task. However it is still seriously sensitive to the changes in dictionary. Thus we expect to have an approach that is more robust if not dependent of a dictionary.

The main contributions of this work are summarized as follows:

- Building semantic textual feature. This approach can capture tags semantic information which is hardly described by the term frequencies models.

- Using *WordNet*-based semantic distance for feature construction. This approach is robust, because this method does not depend on dictionary construction.

4.2 Semantic textual feature using a dictionary

This work presents a novel method for building textual feature defined on semantic distance and describes multi-model approach for Visual Concept Detection and Annotation(VCDA). Nowadays, the tags associated with images have been popularly used in the VCDA task, because they contain valuable information about image content that can hardly be described by low-level visual features. Traditionally the term frequencies model is used to capture this useful text information. However, the shortcoming in the term frequencies model lies in that the valuable semantic information can not be captured. To solve this problem, we propose the semantic bag-of-words(BoW) model which use *WordNet*-based distance to construct the codebook and assign the tags. The advantages of this approach are two-fold: (1) It can capture tags semantic information that is hardly described by the term frequencies model. (2) It solves the high dimensionality issue of the codebook vocabulary construction, reducing the size of the tags representation. The experimental results on the ImageCLEF 2011 show that our approach effectively improves the recognition accuracy.

4.2.1 Our Approach

Textual information coming with an image is basically composed of words. These set of words can be seen as a Bag-of-Words (BoW). The BoW model is a promising textual representation technique for textual categorization. However, the critical limitation of existing BoW model lies in the fact that tags lose semantic information during the dictionary generation process and assignment. Indeed, the BoW kind approaches assume that word terms are basically statistically independent, thereby mismatching tags close in content but with different term vocabulary. Based on the classical approach of BoW, we propose to build a semantic textual feature for an image I by computing the histogram of occurrences of the words of a dictionary in the set of tags associated to this image I. Compared with classical BoW models, semantic BoW is defined as a histogram of textual concepts toward a dictionary where each bin of this histogram represents a concept of the dictionary, whereas its value is the accumulation of the frequency of each word within the tag set toward the underlying concept according to a predefined semantic similarity measure. Meanwhile, in order to reduce the dimensions of feature vector, we tried to group words of the dictionary that have similarity above a certain threshold, using the $Wordnet$ similarity.

4.2.1.1 Semantic BoW feature

Instead of choosing the big frequency words which appear in corpus and assigning the words with the same term vocabulary, we try to employ $Wordnet$ similarity to reduce the size of dictionary and assign the histogram. It can be called componential space model, such as conceptual vector, which describes the meaning of a word by its atoms, its components, attributes, behavior, related ideas, etc. The semantic BoW is defined as a histogram of textual concepts toward a dictionary where each bin of this histogram represents a concept of the dictionary, whereas its value is the accumulation of the frequency of each word within the tag set toward the underlying concept according to a predefined semantic similarity measure. This is in clear contrast to the BoW approaches where the relatedness of textual concepts is simply

Table 4.1: The procedure of the Semantic BoW algorithm

Semantic BoW Model

Input: Training dataset $Tr = \{Tr_1, Tr_2, \ldots, Tr_n\}$ and Testing/Training dataset $Te = \{Te_1, Te_2, \ldots, Te_m\}$.
Output: The K-length feature vector.
Initialization: $frequency\ F$, $threshold\ T_1, T_2$.

- Chosen words which frequencies above F from Tr
- Part of words P=*noun*
- Build clustering on training data(dictionary construction)
  - Construct $W \times W$ matrix M where M_{xy} is the $WordNet$-based distance between W_x and W_y
  - Use matrix to combine the word Where $M_{xy} > T_1$.
  - Dictionary D = $\{d_1, d_2, \ldots, d_K\}$ is constructed
- Build tags representation of Tr and Te data(assignment)
  - For each words $w_i \in Tr_i$ *or* Te_j
  - For each words $w_j \in D$
  - $V_j = V_j + 1$ if compute $SIM(w_i, w_j) > T_2$

ignored as word terms are statistically counted. The procedure of the semantic BoW algorithm is presented as table 4.1

4.2.1.2 Semantic clustering

Our motivation of using the semantic clustering is to reduce the size of the tags representation and capture more semantic information through this approach. The procedure of proposed method is shown in table 4.1. After preprocessing and stemming, the process of suffix removal to generate word stems, the training data sets begins. These word stems and some tags that are popular on internet are used in annotation. But maybe these words are not contained in $WordNet$. In order to avoid discarding these useful information, the tags which appear P times are chosen as the training set. A set of N words is generated. The distance matrix mAPs two $WordNet$ senses to a real number between 0 and 1. A matrix of the pairwise distance is constructed using the metric specified in equation 4.1. A cluster is then

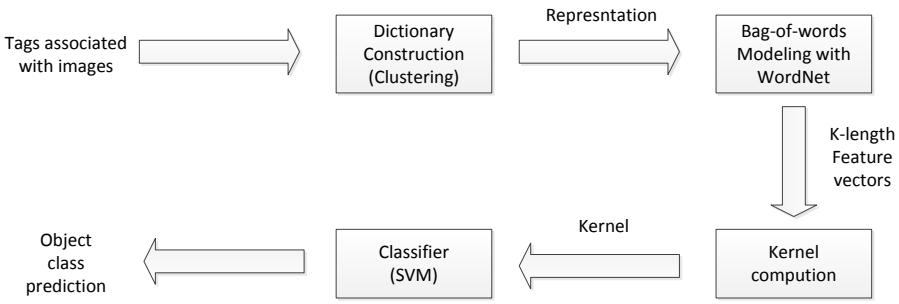


Figure 4.2: The framework of Semantic Bag-of-Words feature.

obtained on the matrix as the threshold T, with the specified target cluster count. The output is a clustering model that can assign an unseen word to a cluster based on its distance to the training word.

4.2.2 The Framework of Experiment

Our framework for VCDA is depicted in Fig 4.2.

4.2.2.1 Dataset and Experimental evaluation

We perform Concept Detection and Annotation on the imageCLEF 2011 Photo Annotation Challenge dataset. The ImageCLEF 2011 dataset are employed in our experiment. The training set for annotation task consists of 8000 photos annotated with 99 visual concepts, and the testing set consists of 10000 photos with EXIF data and Flickr user tags. These 99 concepts include the scene categories (indoor, outdoor, landscape, etc.), depicted objects (car, animal, person, etc.), the representation of image content (portrait, graffiti, art, etc.), events (travel, work, etc.) or quality issues (overexposed, underexposed, blurry, etc.). Thus, this task can be solved by following three different approaches[2]:

- Automatic annotation with visual information only.
- Automatic annotation with Flickr user tags (tag enrichment).
- Multi-modal approaches that consider visual information and/or Flickr user tags and/or EXIF information.

[2]http://http://imageclef.org/2011/photo

For evaluation we use mean average precision (mAP)[Yue *et al.* 2007], i.e., for each test category we obtain a precision/recall curve, and then compute its average precision based on the area under this curve. Finally the mean value over all the categories is computed.

4.2.2.2 Kernel and Classifier

Once all the local descriptors are transformed to fixed-length features, the χ^2 distance is computed to measure the similarity between each pair of the feature vectors F and F' (n is the size of the feature vector):

$$dist_{\chi^2}(F, F') = \sum_{i=1}^{n} \frac{(F_i - F'_i)^2}{F_i + F'_i} \tag{4.3}$$

Then, the kernel function based on this distance is used for SVM to train the classifier:

$$K_{\chi^2}(F, F') = e^{-\frac{1}{D} dist_{\chi^2}(F,F')} \tag{4.4}$$

Where D is the parameter for normalizing the distances. Here D is set to the average distance of all the training data. Finally, the kernel matrix is feeded to SVM.

4.2.3 Results: textual models

Table 4.2: Comparison of different textual models, on ImageCLEF 2011

Textual model	dictionary size	mAP (%)
Term Frequency	5154	32.53
TF/IDF	5154	32.41
LDA	2500	31.35
HTC	2000	32.12
our semantic BoW model	5154	34.71
our semantic BoW model(clustering)	4215	34.62

The semantic BoW model is employed to build the textual feature. The words that appear at least 3 times (a minimum of 3 times in the training set) are used

as the dictionary, resulting in a dictionary of 5154 words. In order to evaluate our approach, we compare Term Frequency, TF/IDF, LDA and HTC[Liu *et al.* 2011b] approach with the proposed Semantic BoW approach. The mAP performances are shown in table 4.2. The results indicate that the semantic BoW outperforms other methods with almost 2% in mAP evaluation. The main reason may be that some tags which come from internet users contain rich image content. But the frequency of this word is low and not included in the dictionary. This word is thus discarded. The semantic model can easily capture this semantic information through the $WordNet$-based distance. Meanwhile, in dictionary the words which have the same semantic meaning are combined with $WordNet$-based distance. The size of dictionary is thus reduced.

Table 4.3: Comparison of our textual model with other's on ImageCLEF 2011

Teams(Textual model)	mAP (%)
BPACAD[Daróczy *et al.* 2011]	34.6
IDMT[Nagel *et al.* 2011]	32.6
MLKD[Xioufis *et al.* 2011]	32.6
LIRIS[Liu *et al.* 2011b]	32.1
our semantic BoW model	34.7
our semantic BoW model(clustering)	34.6

Moreover, in order to evaluate the semantic BoW model, we compare our approach with the textual configuration results which are obtained top 4 in ImageCLEF 2011 challenge. The LIRIS's result which we submitted to ImageCLEF 2011 challenge is another approach. It can be seen that the semantic BoW model outperforms all team's results, as is shown in table 4.3.

4.3 Semantic textual feature without dictionary

This work presents a novel approach to build the textual feature which is independent of dictionary construction. Traditionally the term frequencies model is used to capture this useful textual information. However, the shortcoming in the term frequencies model lies in the fact that the performance seriously depends on the dictionary construction and in the fact that the valuable semantic information can

not be captured. To solve this problem, we propose image distance feature based on tags, which measures the distance between two set of tags associated with images. The advantages of this approach are two-fold: (1) It is robust, because our feature construction approach does not depend on dictionary construction. (2) It can capture tags semantic information which is hardly described by the term frequencies model.

4.3.1 Our Approach

4.3.1.1 Image distance feature(IDF) based on tags associated with images

The previous results suggest that reducing the size of the dictionary reduce the discrimination of concept. That is the reason why we try to keep the maximum of the information contained in the tags associated to images by keeping them as they are and not associating them to words in a dictionary. Classifying an image will then be made by direct computation of the similarity with the sets of tags from images in the training set, using the similarity measure of equation 4.1 between the most similar words in each set of tags. Whereas the previous approach with dictionary is able to treat situation where the number of tags is important (a complete text for instance), this new approach, without dictionary, is only possible in situations where the number of tags associated to each image is rather small. Similarity computation will be too time consuming instead. The procedure for our approach is shown in Table 4.4. With this approach, it avoids relying on the construction of a dictionary.

4.3.1.2 Image distance feature construction

Our motivation of building textual feature directly based on $WordNet$ is to capture tags semantic information and eliminate the influence of the dictionary construction. In this work we restrict ourselves to the noun component of $WordNet$ and use only hyponymy and instance hyponymy relations for textual feature construction. After preprocessing and stemming, the process of suffix removal to generate word stems, Data set $D\{I_i, T_i\}$ consists of image I_i and tags set T_i. The weight between I_i and

Table 4.4: The procedure of the Image distance feature building algorithm.

Semantic textual feature

Input: Training dataset $Tr = \{Tr_1, Tr_2, \ldots, Tr_n\}$ and Testing dataset $Te = \{Te_1, Te_2, \ldots, Te_m\}$.
Output: The n-length feature vector $F = \{f_{ij}\}$.

- preprocess the tags by using a stop-words filter.
- Build tags representation of Tr and Te data
  - For each $Te_i \in Te$ or $Tr_i \in Tr$
  - if Te_i or Tr_i has no tags, return $f_{ij} = 0$.
  - else Te_i or Tr_i has tags.
    - For each tags set $Tr_j \in Tr$
      - For each words $w_x \in Te_i$ or $w_x \in Tr_i$
      - For each words $w_y \in Tr_j$
      - $f_{ij} = f_{ij} + distance(w_x, w_y)$

I_j are measured by T_i and T_j. We compute the distance between each word of tag set T_i and each word of tag set T_j according to Function 4.1. The overview of the experiment procedure is shown in Table 4.4.

4.3.2 The Framework of Experiment

Our framework for VCDA is depicted in Fig 4.3. In this work, we employ the same Dataset and Experimental evaluation with Semantic textual feature using a dictionary, you can see it in chapter 4.2.2.1. Kernel and Classifier is also same with

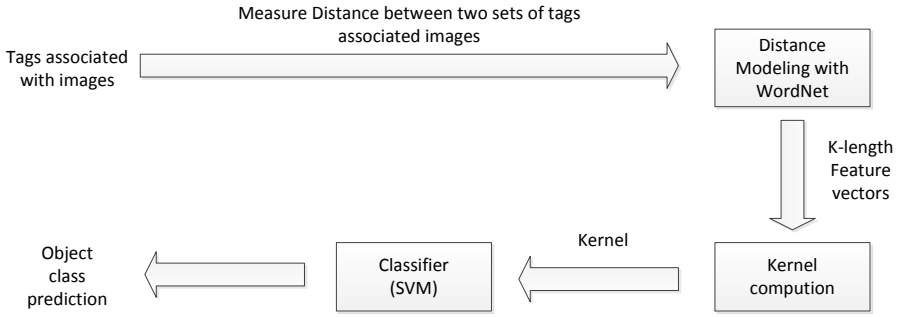


Figure 4.3: The framework of Image distance feature.

chapter 4.2.2.2.

4.3.3 Results: textual models

In the case of ImageCLEF 2011, the average numbers of tags per image is 8.7. Consequently, it is possible to apply this approach. We compare Term Frequency, TF/IDF, LDA and HTC[Liu *et al.* 2011b] approach with the proposed semantic textual feature. The mAP performances are shown in table 4.5. The results indicate that the performance of our textual model is not good , compared with other textual approaches. The main reason may be that our approach only considers the tags semantic relation and nothing about the term frequencies.

Table 4.5: Comparison of different textual models on ImageCLEF 2011.

Textual model	dictionary size	mAP (%)
Term Frequency	5154	32.53
tf/idf	5154	32.41
LDA	2500	31.35
HTC	2000	32.12
Image distance feature(IDF)	-	27.15

Finally, in order to evaluate our approach, we compare our approach with the textual configuration results which are obtained top 4 in ImageCLEF 2011 challenge, as is shown in table 4.6. The LIRIS's result which we submitted to ImageCLEF 2011 challenge is another approach.

Table 4.6: Comparison of our textual model with other's on ImageCLEF 2011.

Teams(Textual model)	mAP (%)
BPACAD	34.6
IDMT[Nagel *et al.* 2011]	32.6
MLKD[Xioufis *et al.* 2011]	32.6
LIRIS[Liu *et al.* 2011b]	32.1
Image distance feature(IDF)	27.15

As it is, this second approach seems not interesting. We will see in the next chapter that it can be improved to give the best results.

4.4 Conclusion

In this chapter, we focused on the problem of how the tags associated with images can benefit for automatic visual concept detection and annotation. We proposed two novel methods to build textual descriptor based on the semantic distance between the user tags. Firstly, we proposed two methods to associate to images a signature computed from the textual information. The first one uses a dictionary and is able to treat situation where textual information is huge (text associated to images on web pages for instance). The second one does not need any dictionary but is only usable in situations where textual information is reduced to a set of few tags. The main contributions are that the semantic textual feature can easily capture semantic information contained in tags which is hardly described by the term frequencies model. Comprehensive experiments were conducted on the ImageCLEF 2011 dataset. Compared with the other approaches, our approach exhibits good preferences for the first approach. The second approach without dictionary gives bad result as it is. We could have leave thing this but trying to understand why these results are that bad. we conclude that it is because this approach without dictionary does not use the frequency information. We will propose an novel approach in the next chapter that will consistently improves the performance of textual classifiers, especially when the concept training set is small.

Chapter 5

Visual Concept Detection and Annotation via Multiple Kernel Learning of multiple models

Contents

5.1 Introduction

There has been a growing demand for image and video data in applications due to the significant improvement in the processing technology, network subsystems and

availability of large storage systems. This demand for visual data has spurred a significant interest in the research community to develop methods to archive, query and retrieve this data based on their content. For this purpose, many general purpose image retrieval systems have been developed to conquer this challenging. There are three frameworks: textual models, visual models, and multimodel approach.

Textual models can be tracked back to 1970s. In early times, the images are manually annotated by text descriptors, which are then used by a database management system to perform image retrieval. There are two disadvantages with this approach. The first is that a considerable level of human labour is required for manual annotation. The second is the annotation inaccuracy due to the subjectivity of human perception. To overcome the above disadvantages in text models, machine learning approaches was introduced, which employ a binary classifier to learn from labeled images. Now that the increasing amount of images which are weak forms of annotation are currently available on the web, there has been considerable interest in the computer vision community to leverage this data to learn recognition models.

In contrast to textual models, visual models was introduced in the early 1980s. In visual models, images are indexed by their visual content, such as color, texture, shapes. A pioneering work was published by Chang in 1984, in which the author presented a picture indexing and abstraction approach for pictorial database retrieval The pictorial database consists of picture objects and picture relations.

The fundamental difference between visual models and textual models retrieval systems is that the human interaction is an indispensable part of the latter system. Humans tend to use high-level features (concepts), such as keywords, text descriptors, to interpret images and measure their similarity. The features automatically extracted using computer vision techniques are mostly low-level features (color, texture, shape, spatial layout, etc.). In general, there is no direct link between the high-level concepts and the low-level features. It has been proven that using only the textual models or visual models is not sufficient for accurate classification.

Thus, the multimodel approach was proposed to automatically predict the visual concepts of images through an effective fusion of textual features along with the visual ones. In contrast to single model, the multimodel approach needs the

fusion of different sources of information for tacking a decision. This fusion can be made using different strategies. An early fusion considers a single feature set including all features extracted from every source of information. This method has the advantage of emphasizing individual features weekly presents in each source. It must however deal with a greater and heterogeneous set of individual features. On the other side, a late fusion combines into one final decision those (decisions) taken individually from each source. Between these two strategies, a lot of intermediate strategies are conceivable. They consist in generating intermediate states (or classes) from different sources and to take a decision based on these intermediate states, using for example hidden markov models, Fuzzy logic or neural networks (and some combination of theses techniques such as neuro-fuzzy models).

This work of efficient strategies can be done according to three axes:

- Description of each modality content: from a light description to a deeper description. In this work, we do not only make use of low-level features including color, shape, texture, SIFT, but also consider higher level textual features such as the classical Bog-of-words approach, LDA. Moreover, we propose the Semantic Bag-of-Words feature, Robust Semantic Bag-of-Words feature. These semantic textual descriptions could lead to a semantic classification of this image according to an ontology tree: landscape, city, indoor/outdoor.

- Integration strategy of each modality content: from an early strategy where content features of each modality are simply integrated into a single feature vector before training and classification, to a late strategy where a decision of partial classification has already been taken for each modality before the final decision. In this work, we consider to combine multiple feature channels for the purpose of efficient image classification in kernel level.

- Supervised training algorithms: a lot of supervised techniques have been developed in the literature, from neural networks to SVM and other data mining methods such as decision trees or instance based learning. However, discriminative kernel based methods, such as SVMs, have been shown to be quite effective for image classification. To use these methods with several feature

Flickr user tags: agra uttar pradesh india taj mahal tajmahal monument mumtazmahal Wonder mughal mughalarchiture muslim muslimart unesco

Flickr user tags: love stupid couple silhouette tree sunset sunrise hulhumale happy coke shade friends art beach blue

Flickr user tags: enero flor flower fleur flores flowers fleurs loveartflowers

Flickr user tags: tucson flower om

Flickr user tags: wordbk littlemars japan tokyo bored creative

Flickr user tags: old abandoned bridge coveredbridge georgia

Figure 5.1: Example images with sparse Flickr user tags.

channels, one needs to combine base kernels computed from them. Multiple kernel learning is an effective method for combining the base kernels. In this work, we consider to employ MKL to combine the multiple feature channels.

5.2 Textual Models and Visual Models

There exist abundant captioned images on the Internet. A textual tags for a given image is very sparse, for instance only counting 8.7 tags in average for the MIR FLICKR collections, as illustrated in Figure 5.1. In this section, we firstly describe descriptor of textual content. Then the visual models are described.

5.2.1 Textual Models

The tags associated with images provides valuable information, which can hardly be described by low-level visual features. In order to make use of these efficient tags, textual models is employed to capture them. According to the vector space model as a vector of terms, each component is a kind of word count of term frequency as exemplified by tf, tf/idf(term frequency inverse document frequency), this model has undergone several extensions, such as LSA, pLSA, LDA, etc. However, image tags are generally sparse text, only having in average 8.7 tags per image. They do not provide enough text content to correctly train frequency extension model. In order to address this drawback, $Wordnet$ is employed to capture the relatedness of semantic concepts, which are introduced in chapter 4. Meanwhile we use $WordNet$-based semantic distance for dictionary construction to reduce the size of the tags representation.

5.2.1.1 textual feature

We have seen that the dominant "bag-of-words" approach falls short to describe the fineness and the relatedness of semantic concepts. Indeed, the BoW kind approaches assume that word terms are basically statistically independent, thereby mismatching text documents close in content but with different term vocabulary. In contrast, we propose Semantic Bag-of-Words feature and Image Distance feature based on tags

associated with images to capture the semantic relatedness of concepts. In this work, we do not only make use of frequency textual features, but also to consider higher level concept textual features. Table 5.1 summarizes all the textual features that we have implemented for the purpose of VCDA.

Table 5.1: Summary of textual features used in our experiments.

Category	Short name	dictionary	Short Description
Frequency	tf	5154	This operator is obtained by computing the number of times a term occurs in a document.
	tf/idf	5154	tf/idf is the product of two statistics, term frequency and inverse document frequency. tf is the frequency of a term in a document. idf is a measure of whether the term is common or rare across all documents.
Semantic	HTC	2000	The HTC is a a histogram of textual concepts toward a dictionary where each bin of this histogram represents a concept of the dictionary, whereas its value is the accumulation of the contribution of each word within the text document toward the underlying concept according to a predefined semantic similarity measure.
	sBoW	5154	sBoW is a histogram of textual concepts toward a dictionary where each bin of this histogram represents a concept of the dictionary, whereas its value is the frequency accumulation of each word within the text document based on semantic similarity measure.
	IDF	–	IMF is a histogram of image distance based on the tags similarity associated with images. whereas its value of each bin of this histogram is the accumulation of the contribution between two sets of tags associated with images.

5.2.2 Visual Models

5.2.2.1 Visual features

The visual appearance of an object has a strong dependency on the viewpoint under which it is recorded. Salient point methods introduce robustness against viewpoint changes by selecting points[van de Sande *et al.* 2010], which can be recovered under different perspectives. Another simpler solution is to use many points, which is achieved by dense sampling. Dense sampling has been shown to be advantageous for scene type classification, since salient points do not capture the entire appearance of an image. For object classification, salient points can be advantageous because they ignore homogenous areas in the image. If the object background is not highly textured, then most salient points will be located on the object or the object boundary.

Commonly the visual content of an image is described by visual descriptors such as color, texture, shape, etc. within a global or a bag of local features. In this work, we make use of several popular local descriptors,including C-SIFT, Rgb-SIFT, Hsv-SIFT,Oppo-SIFT and DAISY, extracted from a dense grid. An image is then modeled as bag-of-visual words using a dictionary of 4000 visual words and hard assignment. Meanwhile, in order to capture the global ambiance and layout of an image, we further compute a set of global features, including descriptions of color information, in terms of LBP, Color LBP [Zhu *et al.* 2010], Table 5.2 summarizes all the visual features that we have implemented for the purpose of VCDA.

5.2.2.2 Bag-of-Features representation

After local feature extraction, each input image is represented by a set of local descriptors. Because of the large number of sampling points (normally more than thousands), it is unreasonable to feed them directly into the classifier. Meanwhile these descriptors can not directly bridge the gap between visual descriptors and the semantic content of image. Therefore, we employ the dominant Bag-of-Features (BoF) method which views an image as an unordered distribution of local image features extracted from dense image points[Mikolajczyk & Schmid 2001] and transform

these high dimensional descriptors to more compact and informative representations. We apply the popular Bag-of-Features (BoF) method here because of its great success in object recognition tasks.

Table 5.2: Summary of visual features using in experiment.

Category	Short name	codebook	Short Description
Global	Hsv-LBP	1311	This operator is obtained by computing LBP for the Hue channel of the HSV color space.
	Inv-LBP	1311	This operator is obtained by computing LBP over all three channels of the transformed color space.
	RGB-LBP	1311	This operator is obtained by computing LBP over all three channels of the RGB color space independently, and then concatenating the results together.
	Oppo-LBP	1311	This operator is obtained by computing LBP over three channels of the opponent color space.
Local	C-SIFT	4000	The C-SIFT feature uses the C invariant. which can be intuitively seen as the gradient (or derivative) for the normalized opponent color space $O1/I$ and $O2/I$.
	RGB-SIFT	4000	For the RGB-SIFT, the SIFT feature is computed for each RGB channel independently.
	HSV-SIFT	4000	This operator is obtained by computing HSV-SIFT for the Hue channel of the HSV color space..
	Oppo-SIFT	4000	Oppo-SIFT describes all the channels in the opponent color space using SIFT features.
	DAISY	4000	DAISY descriptor computed on a dense grid DAISY descriptor computed on a dense descriptor computed on a dense grid.

The main idea of the BoF is to represent an image as an unorderless collection of local descriptors. More precisely, a visual vocabulary is constructed at first by applying a clustering algorithm such as k-means on the training data, and each cluster center is considered as a 'visual word' in the vocabulary. All feature descrip-

tors extracted from an image are then quantized to their closest 'visual word' in an appropriate metric space. Finally the images are represented as fix-length vectors. The number of feature descriptors assigned to each 'visual word' is then accounted into a histogram as the final BoW representation. Since the BoW modeling ignores all spatial information of local features, we also consider spatial pyramid to take into account coarse spatial relationship between them.

5.3 Multiple Kernels Learning

Due to the possibly large intraclass feature variations, using only a single unified kernel-based classifier may not satisfactorily solve the problem. Instead of selecting a single kernel, MKL learns a convex kernel combination and the associated classifier simultaneously; the combination of multi-kernels is defined as follows:

$$K(x_i, x) = \sum_{m=1}^{M} d_m K_m(x_i, x) \tag{5.1}$$

with $\sum_{m=1}^{M} d_m = 1$ and $d_m \geq 0 \quad \forall m$ where M is the total number of kernels, $K_m = \phi_m(x_i)\phi_m(x_j)$ is a positive definite kernel which represents the dot product in feature space ϕ, and $\{d_m\}_{m=1}^{M}$ are kernel weights which are optimized during training. Each K_m can employ different kernel functions and use different feature subsets or data representations.

For binary classification, given the learning set $\{x_i, y_i\}_{i=1}^{M}$, where x_i belongs to some input data and y_i is the label of x_i, the decision function of canonical MKL is given as follows:

$$f(x) = \sum_{i=1}^{N} \alpha_i^* y_i \sum_{m=1}^{M} d_m K_m(x_i, x) + b^* \tag{5.2}$$

Where $\{\alpha_i^*\}_{i=1}^{N}$ and b^* are the coefficients of the classifier, corresponding to the lagrange multipliers and the bias in the canonical SVM problem. To solve the MKL problem efficiently, the SMO-MKL algorithm is used to optimise the l_p MKL dual[Vishwanathan *et al.* 2010].

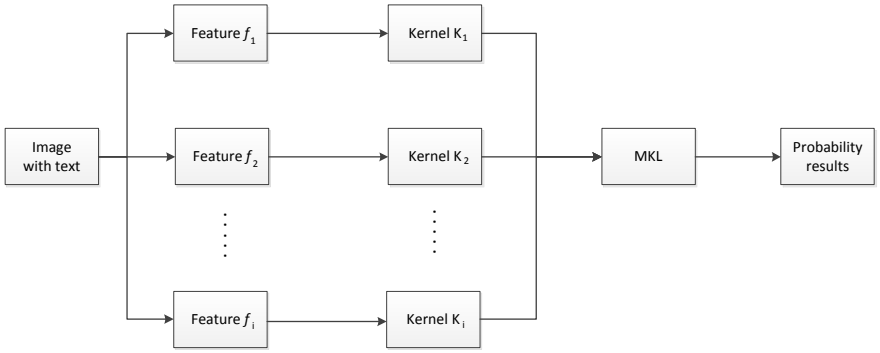


Figure 5.2: The framework of multimodel approach.

The primal can therefore be formulated as

$$\begin{aligned} min \quad & \sum_k \frac{1}{d_m} w_k w_k^T + C \sum_i \xi_i \\ s.t. \quad & y_i \sum_k \phi_k(x_i) + y_i b \geq 1 - \xi_i \quad \forall i \\ & \xi_i \geq 0 \quad \forall i \\ & \sum_m d_m = 1, \quad d_m \geq 0 \quad \forall m \end{aligned} \tag{5.3}$$

where b is the bias, ξ_i is the slack afforded to each data point and C is the regularization parameter. The solution to the above MKL formulation is based on a gradient descent on the SVM objective value. An iterative method alternates between determining the SVM model parameters using a standard SVM solver and determining the kernel combination weights using a projected gradient descent method.

5.4 The Approach for VCDA

Our framework for VCDA is depicted in Fig 5.2.

5.4.1 Fusion and Classification

The chi-square kernel(χ^2 distance) is used to measure the similarity between two feature vectors F and F' (n is the size of the feature vector). Then, the kernel function based on this distance is used for MKL to train the classifier:

$$K_{\chi^2}(F, F') = e^{-\frac{1}{D}\sum_{i=1}^{n}\frac{(F_i - F'_i)^2}{F_i + F'_i}} \tag{5.4}$$

Where D is the parameter for normalizing the distances. Here D is set to the average distance of all the training data. Finally features are presented as kernel matrixes. Kernels with different components or building approaches usually capture different and complementary content information of image, making them have different discriminative power with different weights. Once giving kernels, MKL seeks to the best combination-weights of these kernels.

5.4.2 Data set and Experimental evaluation

In our experiment the ImageCLEF 2011 dataset with 99 concepts are employed. The training set consists of 8000 photos, and the testing set consists of 10000 photos. All photos are associated with EXIF data and Flickr user tags, These 99 concepts include the scene categories, depicted objects, the representation of image content, events or quality issues. For evaluation, we use mean average precision (mAP)[Yue *et al.* 2007].

Thus, this task can be solved by following three different approaches[1]:

- Automatic annotation with visual information only.
- Automatic annotation with Flickr user tags (tag enrichment).
- Multi-modal approaches that consider visual information and/or Flickr user tags and/or EXIF information.

i.e., for each test category we obtain a precision/recall curve, and then compute its average precision based on the area under this curve. Finally the mean value over all the categories is computed.

[1]http://http://imageclef.org/2011/photo

5.5 Experimental Evaluation

The MKL approach is employed to fuse the textual model and the visual model. The different types of visual models are fused with textual models.

5.5.1 Results: fusion of textual models

The first association we made using MKL was dedicated to the improvement of our IDF method of chapter 4. Fusion was made with tf/idf classical feature on words of the tags set of an image. The results we obtain are presented on table 5.3 under the name of tf/idf_IDF(MKL), which fuses tf/idf and IDF with the MKL approach. The results are really convincing and outperform the other classical textual descriptors.

Table 5.3: Comparison of different textual models on ImageCLEF 2011 dataset.

Textual model	dictionary size	mAP (%)
Term Frequency	5154	32.53
tf/idf	5154	32.41
LDA	2500	31.35
HTC	2000	32.12
semantic BoW model	5154	34.71
semantic BoW model(clustering)	4215	34.62
Image distance feature(IDF)	-	27.15
tf/idf_IDF(MKL)	-	37.48

Table 5.4: Comparison of our textual fusion results with other's on ImageCLEF 2011.

Teams(Visual model)	mAP (%)
BPACAD[Daróczy *et al.* 2011]	34.6
IDMT[Nagel *et al.* 2011]	32.6
MLKD[Xioufis *et al.* 2011]	32.6
LIRIS[Liu *et al.* 2011b]	32.1
tf/idf_IDF(MKL)	37.5

Moreover, in order to evaluate the tf/idf_IDF(MKL) model, we compare our approach with the textual configuration results which are obtained top 4 in ImageCLEF 2011 challenge. The LIRIS's result which we submitted to ImageCLEF

2011 challenge is another approach. It can be seen that tf/idf_IDF(MKL) model outperforms all team's purely textual results, as is shown in Table 5.4.

5.5.2 Results: fusion of visual models

We apply different types of visual features to build the visual models and fuse same types of visual features with MKL respectively on ImageClEF 2011 dataset. The experimental results of each single visual feature and fusion approach are shown in Fig.5.3. For each single visual feature, we can see that the color SIFT based features outperform other descriptors. The performances of color SIFT features obtain about $30\% \sim 34\%$ mAP value. Moreover compared with single visual feature, the performance of multi-visual model is better.

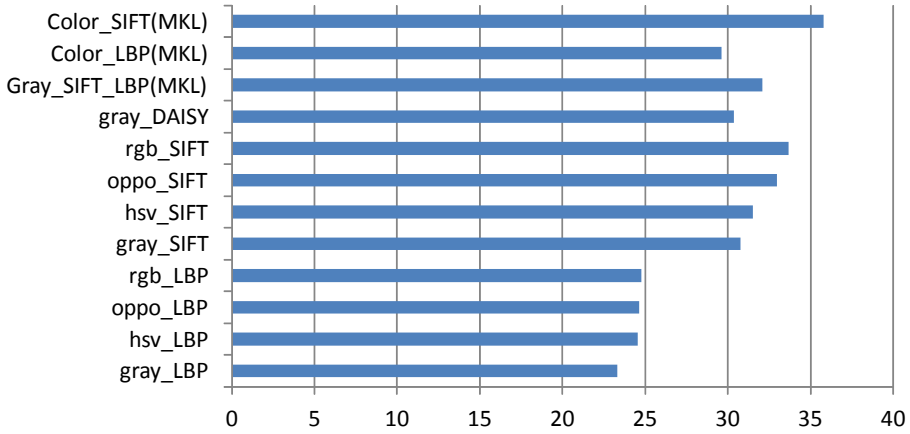


Figure 5.3: The mAP performance of different visual models.

Table 5.5 shows the performance of different teams who participated the ImageCLEF 2011 challenge. TUBFI's, CAEN's, ISIS's and BPACAD's purely visual model ranked the $1^{st}, 2^{nd}, 3^{rd}$ and 4^{th}. Compared with their purely visual results , the performance of our visual model is comparable.

5.5.3 Results: fusion of visual models and textual models

5.5.3.1 Results: fusion of visual models and semantic BoW models

The MKL approach is employed to fuse semantic BoW model and the visual model. In order to evaluate our approach, We investigated the results of TUBFI, Liris,

Table 5.5: Comparison of our visual model with other's on ImageCLEF 2011.

Teams(Visual model)	mAP (%)
TUBFI[Binder *et al.* 2011]	38.8
CAEN[Su & Jurie 2011]	38.2
ISIS[van de Sande & Snoek 2011]	37.5
BPACAD[Daróczy *et al.* 2011]	36.7
Color_LBP_SIFT(MKL)	37.4

BPACAD, ISIS and MLKD, whose top 5 multimodel approaches ranked in the challenge 2011 on mAP evaluation, as shown in table 5.6. TUBFI applied non-sparse multiple kernel learning and multi-task learning to build classifiers. To build the textual features, they used BoW and Markov random walks based on the Flickr user tags. Compared with other team's results, our approach gets the best result of 45.33% mAP.

Table 5.6: Comparison of different multimodel approach on ImageCLEF 2011

Multi model	mAP (%)
TUBFI	44.3
LIRIS	43.7
BPACAD	43.6
ISIS	43.3
MLKD	40.2
sBoW_visual(MKL)	45.3

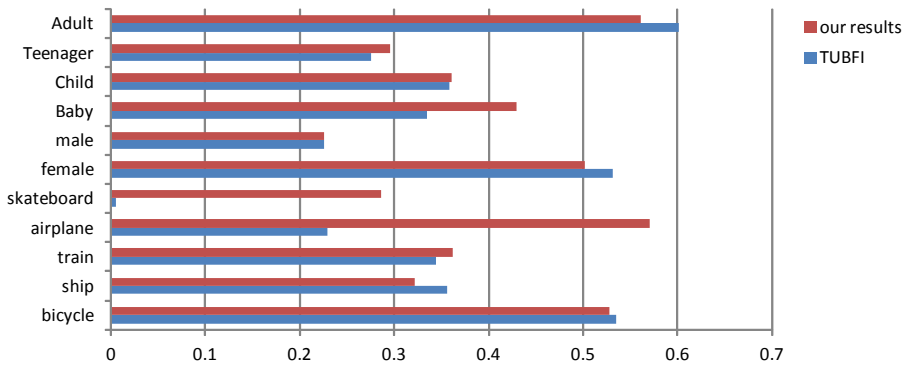


Figure 5.4: A part of the Average Precision per concept of our sBoW_visual multimodel runs compared to TUBFI's.

Fig 5.4 shows the Average Precision per concept in detail, and it can be noticed

that our results significantly outperform the TUBFI's best run on the concepts of airplane and skateboard. Analysis shows that the number of training samples for these concepts are only 41 and 12, which makes it extremely difficult to classify those concepts. However, our textual features improve the performance of our visual classifiers regarding to these cases.

5.5.3.2 Results: fusion of visual models and IDF models

The MKL approach is employed to fuse the textual model and the visual model. The different types of visual models are fused with textual models. The experimental results are shown in Table 5.7. The results notices that combining multiple feature channels can improve the performances. Meanwhile we investigated the results of TUBFI, Liris, BPACAD, ISIS and MLKD, whose multimodel approaches ranked in top 5 of the challenge 2011 on mAP evaluation, as shown in Table 5.8. TUBFI applied non-sparse multiple kernel learning and multi-task learning to build classifiers. To build the textual features, they used BoW and Markov random walks based on the Flickr user tags. Compared with other team's results, our approach gets the best result of 45.73% mAP.

Table 5.7: The mAP performance of different multimodel approach on ImageCLEF 2011.

Multi model(MKL)	mAP (%)
LBP_tf/idf_IDF	42.26
SIFT_tf/idf_IDF	44.24
LBP_SIFT_tf/idf_IDF(MKL)	45.73

Table 5.8: Comparison of our multimodel with other's on ImageCLEF 2011.

Teams(multimodel)	mAP (%)
TUBFI	44.3
LIRIS	43.7
BPACAD	43.6
ISIS	43.3
MLKD	40.2
LBP_SIFT_tf/idf_IDF(MKL)	45.7

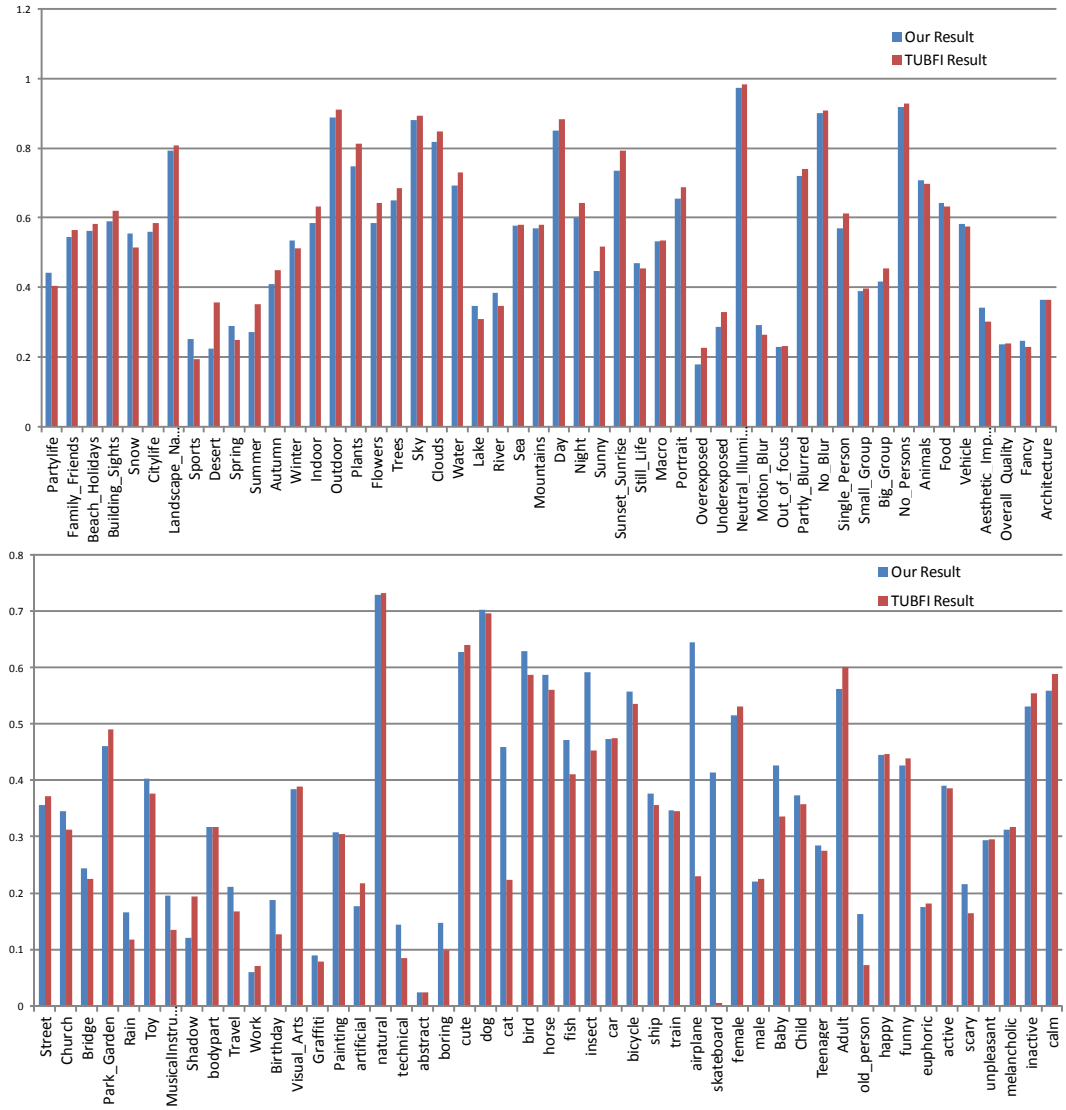


Figure 5.5: The Average Precision per concept of our best multimodel runs compared to TUBFI's.

Fig 5.5 shows the Average Precision per concept in detail, and it can be noticed that our results significantly outperform the TUBFI's best run on the concepts of airplane, skateboard, baby, and cat. Analysis shows that the number of training samples for these concepts are only 41, 12, 90 and 53, which makes it extremely difficult to train classifiers and apply it to classify those concepts. However, our textual IDF can effectively captures tags information and improves the performance of our visual classifiers regarding to these cases.

5.6 Conclusion

In this chapter, we considered a multimodel approach to address the active research topic of the visual concept detection task in images. Our two novel sematic textual features we previous proposed which both use semantic similarity measure based on WordNet are employed to fuse. Firstly, we employ MKL approach to combine our IDF feature with term frequency feature, it is especially interesting because IDF model gives better classification results than the sBoW model when combined with tf/idf model. It means that IDF is more interesting to capture pure semantic concept information. Secondly, we propose a fusion method based on MKL to regroup these multimodal descriptors. The visual features that we used here are very classical features (LBP and SIFT). Other visual features will be tested in future works but the point here was to demonstrate all the interest of our MKL approach of fusion. Finally, in order to take advantages of both textual and visual information, we employ MKL approach to combine our sBoW feature with visual features to get classification results and combine our IDF feature with visual features to get classification results. Our results show that joint use of user textual tags and visual descriptions can better bridge the gap between high level semantic concepts and low-level visual features. As the experimental results, we think that we reach the goal because our results are at the comparable or even better than the ones of the best teams on the international images classification competition ImageCLEF 2011.

Chapter 6

Conclusions and Future Work

Contents

6.1 Conclusions

In this thesis, we considered the multimodal approach to address the active research topic of the visual concept detection task which consists in labeling a real world image according to the concept it contains given a set of categories under consideration. Images came from the real world without any position restriction on the processed images. This means image content may be heterogeneous, ambiguous, and also acquired under poor conditions. Moreover, we have to deal with the problems inherent to image content like the wide variety of shape and appearance of objects inside a category, and due to the representation of a concept in an image, such as various scales and orientations, as well as illumination and occlusion problems. Due to all these difficulties, it is hard to solve the problem by using methods that only depend on the visual information. The text associated with images provides valuable information about image content that can hardly be described by low-level visual features. These abundant textual captions associated with image convey rich semantic meanings and frequency information. Mixing use of user textual tags, and visual descriptions, the multimodal approach can better bridge the gap between high level semantic concepts and low-level visual features.

Firstly, we propose a novel SM-LBP descriptors which can obtain multi-scale

patterns and provide a patch texture representation. Moreover, in order to deal with the deficiency of color information and sensitivity to non-monotonic lighting condition changes, SMC-LBP descriptor is proposed. The main contributions are that the SM-LBP and SMC-LBP not only have more discriminating power by obtaining more local information, but also possess invariance properties to different lighting condition changes. In addition, they keep the advantage of computational simplicity from the original LBP descriptor. The proposed descriptors are validated by applying on the PASCAL VOC 2007 image benchmark. Compared with the original LBP, the experimental results exhibit better recognition accuracy.

Secondly, we introduced a novel approach to use local binary descriptors for the task of VOC. The main contributions are proposing a new encoding method to address the high dimensionality issue of the traditional binary bitstring encoding, and to adopt Hamming distance with the BoF model for visual vocabulary construction and histogram assignment. HD is suitable for computer instruction because it performs an XOR operation. In contrast to other distances, HD spends less time and needs less computer resource. The proposed approach was validated by applying on the LBP feature on the PASCAL VOC 2007 dataset. Compared with the original LBP, it exhibited better recognition accuracy. Meanwhile, we extended the LBP to multi-scale form by directly concatenating binary bitstrings, and also obtained a better performance than the traditional multi-scale fusion in histogram level. The time consumption is very reasonable.

Thirdly, we focused on the problem of how the tags associated with images can benefit for automatic visual concept detection and annotation. We proposed two novel methods to build textual descriptor based on the semantic distance between the user tags. The first one uses a dictionary and is able to treat the situation where textual information is huge (text associated with images on web pages for instance). The second one does not need any dictionary, but is only usable in situations where textual information is reduced to a set of few tags. The main contributions are that the semantic textual feature can easily capture semantic information contained in tags which is hardly described by the term frequency model. Comprehensive experiments were conducted on the ImageCLEF 2011 dataset. Compared with the

other approaches, our approach exhibits good preferences. From the experimental results, we conclude the following: Based on the proposed approach, it consistently improves the performance of visual classifiers, especially when the concept training set is small.

Finally, we considered a multimodal approach to address the active research topic of the visual concept detection task in images. Our results show that joint use of user textual tags, and visual descriptions can better bridge the gap between high level semantic concepts and low-level visual features. However, it is especially interesting because it gives better classification results than the sBoW model when combined with term frequency analysis. Both use semantic similarity measure based on WordNet. Finally, in order to take advantages of both textual and visual information, we propose a fusion method based on MKL to regroup these multi-modal descriptors. The visual features that we used here are very classical features (LBP and SIFT). Other visual features will be tested in future works, but the point here was to demonstrate all the interest of our MKL approach of fusion. We think that we reach the goal because our results are at the comparable or even better than the ones of the best teams in the international images classification competition ImageCLEF 2011.

6.2 Perspectives for Future Work

We present in this section some perspectives for future research directions.

Compared with the original LBP and other texture descriptors, our SM-LBP descriptors exhibits the better recognition accuracy in the task of VOC. However these novel texture descriptors are still not used in VCDA task. In future work, we will consider to employ these novel texture descriptors to combine with other kinds of features.

Our proposed BoF model with Hamming distance was validated by the LBP feature of PASCAL VOC 2007 dataset. Future work could consider to use other local binary descriptors (e.g. BRIEF) in our framework for the task of VOC as well as texture classification. Moreover, the proposed approach can be extended to dif-

ferent color local binary descriptors(e.g. HSV-BRIEF and OPPONENT-BRIEF) to improve the performance. In addition, other local binary descriptors (e.g. BRIEF) will be considered to be modeled by this novel BoF model with Hamming distance. Thus these local binary descriptors will be effectively use in Multimodel approach.

For our proposed textual features, our semantic BoW feature will extend to apply to perform an efficient classification of multimedia documents as they are found on web. Because this novel feature is not only capture the fineness and the relatedness of semantic concepts, it also can be speedily computed.

In order to integrate different kinds of modality content from different sources, we employ MKL to fuse basis kernel with different parameter configurations. Future work, how to effectively fuse different kinds of models also remains a problem, while hidden markov models[Chu & Huang 2007], Fuzzy logic[Nedeljkovic 2004] or neural networks[Schmidhuber 2012] provide some ideas.

Appendix A

A Participation in the Popular Challenges

Contents

During this thesis, we participate to the popular challenge in computer vision community: Photo Annotation of ImageCLEF 2011[1], Photo Annotation and Retrieval of ImageCLEF 2012[2], partly based on the work of this thesis.

A.1 Participation in Photo Annotation of ImageCLEF 2011

Photo Annotation of ImageCLEF 2011 challenge is a popular benchmark for the visual concept detection and annotation. This task is a multi-label classification challenge. It aims at the automatic annotation of a large number of consumer photos with multiple annotations. A detailed introduction of the ImageCLEF can be found in charter 2.5.2.

In 2011, the training set for annotation task consists of 8000 photos annotated with 99 visual concepts, and the testing set consists of 10000 photos with EXIF data and Flickr user tags. These 99 concepts include the scene categories (indoor,

[1] http://www.imageclef.org/2011/photo
[2] http://www.imageclef.org/2012/photo

outdoor, landscape, etc.), depicted objects (car, animal, person, etc.), the representation of image content (portrait, graffiti, art, etc.), events (travel, work, etc.) or quality issues (overexposed, underexposed, blurry, etc.). the challenge has provided multimodel approaches that consider visual information and/or Flickr user tags and/or EXIF information.

For this task, we firstly propose two kinds of textual features to extract semantic meanings from text associated to images: one is based on semantic distance matrix between the text and a semantic dictionary, and the other one carries the valence and arousal meanings by making use of the Affective Norms for English Words (ANEW) dataset. Meanwhile, we investigate efficiency of different visual features including color, texture, shape, high level features, and we test four fusion methods to combine various features to improve the performance including min, max, mean and score.

On one hand, based on previous two kinds of textual features methods proposed, we build 10 textual features on different words semantic distance and dictionary(dict119 and dict1034). On the other hand, we extracted from each image the dense SIFT descriptor and a set of global features, including Color Histogram, Color Moments, Color Coherence Vectors, Gray Level Co-occurrence Matrix, Local Binary Patterns, Edge Histogram, and Line Segment, to describe the visual content of images. A vocabulary of 4000 visual words was created for the Bag-of-Features model of SIFT, and hard assignment was adapted to build the histogram. The SVM classifier was used for classification, and the Chi-square distance was computed as the kernel of SVM for all kinds of features. Finally, we perform fusion methods including min, max, mean, score(mAP as the score), and selected best fusion among 4 methods(min, max, mean, score) for each concept.

We performed our runs based on following configuration:

- **textual model** we selected top 4 features among 10 textual features for each concept according to mAP, and use the mAP as score to combine the output of probability measurements of classifiers. We selected the threshold based on distribution of the training set.

- **textual+visual model** we selected top 21 features among 34 visual and textual features for each concept according to mAP, and use the mAP as score to combine the output of probability measurements of classifiers. We selected the threshold based on best F-measure on validation set.

- **textual model** we selected top 5 features among 10 textual features for each concept according to mAP, and use the mAP as score to combine the output of probability measurements of classifiers. We selected the threshold based on best F-measure[Sang & Meulder 2003] on validation set.

- **visual model** we selected top 5 features among 24 visual features for each concept according to mAP, and use the mAP as score to combine the output of probability measurements of classifiers. We selected the threshold based on best F-measure on validation set.

- **textual+visual model** we selected top 22 features among 24 visual and textual features for each concept according to mAP, and use the mAP as score to combine the output of probability measurements of classifiers. We selected the threshold based on distribution of the training set.

In this year, we submitted 5 runs based on above configuration and features, and among the 5 runs, the 5^{th} one achieved the best performance, which indicated that the combination of textural and visual features outperform than the other runs. The runs are evaluated by three measures to determine the quality of the annotations. One for the evaluation per concept and two for the evaluation per photo. The evaluation per concept was performed with the Mean interpolated Average Precision(mAP). The evaluation per example was performed with the example-based F-Measure(F-ex) and the Semantic R-Precision(SR-Precision)[Euzenat 2007]. The results is shown in Table A.1

Compared our best result with other team's best result of 5 runs submitted, we achieved mAP (Mean Average Precision) of 45.3%, and ranked 2/18 by teams, as shown in Table A.2.

Table A.1: The results of our submitted runs ImageCLEF 2012.

Submitted runs	mAP (%)	F-ex (%)	SR-Precision (%)
text model 1	31.76	43.17	67.49
visual text model 2	42.96	57.57	71.74
text model 3	32.12	40.97	67.57
visual model 4	35.54	53.94	72.50
visual text model 5	43.69	56.69	71.82

Table A.2: Comparison of our results with other's teams on ImageCLEF 2011.

Teams(multimodel)	mAP (%)
TUBFI	44.3
LIRIS	43.7
BPACAD	43.6
ISIS	43.3
MLKD	40.2

A.2 Participation in Photo Annotation of ImageCLEF 2012

In 2012, to improve the performance of our recognition system, we have proposed the Histogram of Textual Concepts (HTC) textual feature to capture the relatedness of semantic concepts. In contrast to term frequency-based text representations mostly used for visual concept detection and annotation, HTC relies on the semantic similarity between the user tags and a concept dictionary. Moreover, a Selective Weighted Late Fusion (SWLF) is introduced to combine multiple sources of information which by iteratively selecting and weighting the best features for each concept at hand to be classified. The results have shown that the combination of our HTC feature with visual features through SWLF can improve the performance significantly.

We submitted 5 runs to the ImageCLEF 2012 photo annotation challenge (2 textual model, 1 visual model and 2 multimodal models). we performed our runs based on the following configuration:

- **textual model** the combination of the top 4 features among the 11 textual features for each concept based on the weighted score SWFL scheme.

- **textual model** the combination of the top 6 features among the 11 textual features for each concept based on the weighted score SWFL scheme.
- **visual model** the combination of the top 5 features among the 24 visual features for each concept based on the weighted score SWFL scheme.
- **multimodal model**the combination of the top 22 features among the 43 visual and textual features for each concept based on the weighted score SWFL scheme.
- **multimodel model**the combination of the top 26 features among the 43 visual and textual features for each concept based on the weighted score SWFL scheme.

The results obtained by our 5 runs are given in Table A.3. The best performance was provided by our multimodal models which outperformed the purely textual and purely visual ones. Moreover, our best model obtained the first rank based on the MiAP among the 80 runs submitted to the challenge, as is shown in Table A.4. In this year, the Geometric Mean Average Precision(GMAP) is employed. This evaluation measure is an extension to mAP.

Table A.3: The results of our submitted runs on ImageCLEF 2012.

Submitted runs	mAP (%)	GMAP (%)	F-ex (%)
text model 1	33.28	27.71	39.17
text model 2	33.38	27.59	46.91
visual model 3	34.81	28.58	54.37
multimodal model 4	43.66	38.75	57.63
multimodal model 5	43.67	38.77	57.66

Table A.4: Comparison of our results with other's teams on ImageCLEF 2012.

Teams(multimodel)	mAP (%)
LIRIS	43.67
DMS-SZTAKI	42.56
CEA LIST	41.59
ISI	41.36

Appendix B

Publications

During this thesis, 7 papers have been published, including 1 paper in an international journal and 6 papers in international conferences. In addition, 1 conference papers have been submitted for review.

B.1 Accepted Paper in International Journal:

1. Ningning Liu, Emmanuel Dellandrea, Chao Zhu, **Yu Zhang**, Charles-Edmond Bichot, Stephane Bres, Bruno Tellez, Liming Chen " Multimodal Recognition of Visual Concepts using Histograms of Textual Concepts and Selective Weighted Late Fusion Scheme" Computer Vision and Image Understanding(CVIU).

B.2 Accepted Papers in International Conferences:

1. **Yu Zhang**, Stphane Bres, and Liming Chen, "Semantic Bag-of-Words Models for Visual Concept Detection and Annotation", The 8th International Conference on SIGNAL IMAGE TECHNOLOGY and INTERNET BASED SYSTEMS (SITIS 2012).

2. **Yu Zhang**, Stphane Bres, and Liming Chen, " Sampled Multi-scale Color Local Binary Patterns", the 8th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (Visapp 2013).

3. **Yu Zhang**, Chao Zhu, Stphane Bres, and Liming Chen, " Encoding Local Binary Descriptors by Bag-of-Features with Hamming Distance for Visual

Object Categorization" The 35th the annual European Conference on Information Retrieval conference (ECIR 2013).

4. **Yu Zhang**, Stphane Bres, and Liming Chen, "Visual Concept Detection and Annotation via Multiple Kernel Learning of multiple models" The 17th International Conference on Image Analysis and Processing (ICIAP 2013).

5. Ningning Liu, **Yu Zhang**, Emmanuel Dellandrea, Stphane Bres, and Liming Chen, "LIRIS-Imagine at ImageCLEF 2011 Photo Annotation Task". (CLEF 2011).

6. Ningning Liu, Emmanuel Dellandrea, Liming Chen, Aliaksandr Trus, Chao Zhu, **Yu Zhang**, Charles-Edmond Bichot, Stphane Bres, "LIRIS-Imagine at ImageCLEF 2012 Photo Annotation Task" (CLEF 2012).

B.3 Submitted Papers in International Conference:

1. **Yu Zhang**, Stephane Bres, Liming Chen, "Images Classification using Multiple Kernels learning based Fusion of Textual and Visual Features" The 21st International Conference on Image Processing (ICIP 2014).

Bibliography

[Abdel-Hakim & Farag 2006] Alaa E. Abdel-Hakim and Aly A. Farag. *CSIFT: A SIFT Descriptor with Color Invariant Characteristics.* In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2, CVPR '06, pages 1978–1983, Washington, DC, USA, 2006. IEEE Computer Society. 26

[Al-Hamami & Al-Rashdan 2010] Alaa Al-Hamami and Hisham Al-Rashdan. *Improving the Effectiveness of the Color Coherence Vector.* Int. Arab J. Inf. Technol., vol. 7, no. 3, pages 324–332, 2010. 17

[Ayache *et al.* 2007] Stephane Ayache, Georges Quenot and Jerome Gensel. *Classifier Fusion for SVM-Based Multimedia Semantic Indexing.* In ECIR 2007, Rome, apr 2007. 58

[Bay *et al.* 2008] Herbert Bay, Andreas Ess, Tinne Tuytelaars and Luc Van Gool. *Speeded-Up Robust Features (SURF).* Comput. Vis. Image Underst., vol. 110, no. 3, pages 346–359, June 2008. 14, 26, 66

[Binder *et al.* 2011] Alexander Binder, Wojciech Samek, Marius Kloft, Christina Müller, Klaus-Robert Müller and Motoaki Kawanabe. *The Joint Submission of the TU Berlin and Fraunhofer FIRST (TUBFI) to the ImageCLEF2011 Photo Annotation Task.* In Vivien Petras, Pamela Forner and Paul D. Clough, editeurs, CLEF (Notebook Papers/Labs/Workshop), 2011. 120

[Bishop 1995] Christopher M. Bishop. Neural networks for pattern recognition. Oxford University Press, Inc., New York, NY, USA, 1995. 6

[Blas *et al.* 2008] Morten Rufus Blas, Motilal Agrawal, Aravind Sundaresan and Kurt Konolige. *Fast color/texture segmentation for outdoor robots.* In IROS, pages 4078–4085, 2008. 20, 81

[Blei *et al.* 2003] David M. Blei, Andrew Y. Ng and Michael I. Jordan. *Latent dirichlet allocation.* J. Mach. Learn. Res., vol. 3, pages 993–1022, 2003. 6, 44, 47

[Boser *et al.* 1992] Bernhard E. Boser, Isabelle M. Guyon and Vladimir N. Vapnik. *A Training Algorithm for Optimal Margin Classifiers.* In David Haussler, editeur, Proceedings of the 5th Annual Workshop on Computational Learning Theory (COLT'92), pages 144–152, Pittsburgh, PA, USA, July 1992. ACM Press. 55

[Burghouts & Geusebroek 2009] G. J. Burghouts and J. M. Geusebroek. *Performance Evaluation of Local Colour Invariants.* Computer Vision and Image Understanding, vol. 113, pages 48–62, 2009. 26

[Calonder *et al.* 2010] Michael Calonder, Vincent Lepetit, Christoph Strecha and Pascal Fua. *BRIEF: binary robust independent elementary features.* In Proceedings of the 11th European conference on Computer vision: Part IV, ECCV'10, pages 778–792, Berlin, Heidelberg, 2010. Springer-Verlag. 6, 7, 28, 67, 81, 83

[Caputo *et al.* 2005] Barbara Caputo, Eric Hayman and P. Mallikarjuna. *Class-Specific Material Categorisation.* IEEE International Conference on Computer Vision (ICCV), vol. 2, pages 1597–1604, 2005. 73, 88

[Chang & Lin 2011] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: A Library for Support Vector Machines.* ACM Transactions on Intelligent Systems and Technology, vol. 2, pages 27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`. 57, 76

[Chatfield *et al.* 2011] K. Chatfield, V. Lempitsky, A. Vedaldi and A. Zisserman. *The devil is in the details: an evaluation of recent feature encoding methods.* In British Machine Vision Conference (BMVC), 2011. 14, 15, 66, 67, 76

[Chu & Huang 2007] Stephen M. Chu and Thomas S. Huang. *Audio-Visual Speech Fusion Using Coupled Hidden Markov Models.* In CVPR. IEEE Computer Society, 2007. 128

[Cortes & Vapnik 1995] Corinna Cortes and Vladimir Vapnik. *Support-Vector Networks.* In Machine Learning, pages 273–297, 1995. 6, 53

[Csurka *et al.* 2004] G. Csurka, C. Bray, C. Dance and L. Fan. *Visual categorization with bags of keypoints.* Workshop on Statistical Learning in Computer Vision, ECCV, pages 1–22, 2004. 29, 67

[Dalal & Triggs 2005] Navneet Dalal and Bill Triggs. *Histograms of Oriented Gradients for Human Detection.* In Cordelia Schmid, Stefano Soatto and Carlo Tomasi, editeurs, International Conference on Computer Vision & Pattern Recognition, volume 2, pages 886–893, INRIA Rhône-Alpes, ZIRST-655, av. de l'Europe, Montbonnot-38334, June 2005. 28, 67

[Daróczy *et al.* 2011] Bálint Daróczy, Róbert Pethes and András A. Benczúr. *SZTAKI @ ImageCLEF 2011.* In Vivien Petras, Pamela Forner and Paul D. Clough, editeurs, CLEF (Notebook Papers/Labs/Workshop), 2011. 102, 118, 120

[Daugman 1988] John G. Daugman. *Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression.* volume 36, pages 1169–1179, 1988. 19

[Deerwester *et al.* 1990] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer and Richard Harshman. *Indexing by latent semantic analysis.* JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE, vol. 41, no. 6, pages 391–407, 1990. 44, 45

[Dumais 1991] Susan T. Dumais. *Improving the retrieval of information from external sources.* Behavior Research Methods, Instruments, & Computers, vol. 23, no. 2, pages 229–236, 1991. 40

[Euzenat 2007] Jérôme Euzenat. *Semantic Precision and Recall for Ontology Alignment Evaluation.* In Proceedings of the 20th International Joint Conference on Artifical Intelligence, IJCAI'07, pages 348–353, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc. 131

[Fellbaum 1998] Christiane Fellbaum, editeur. Wordnet an electronic lexical database. The MIT Press, Cambridge, MA ; London, May 1998. 9, 95, 96

[Finlayson *et al.* 2005] Graham D. Finlayson, Steven D. Hordley and Ruixia Xu. *Convex programming colour constancy with a diagonal-offset model.* In ICIP (3)'05, pages 948–951, 2005. 86

[Fletcher 2008] Tristan Fletcher. *Support Vector Machines Explained.* 2008. 54, 55

[Gaasieniec *et al.* 2000] Leszek Gaasieniec, Jesper Jansson and Andrzej Lingas. *Approximation Algorithms for Hamming Clustering Problems.* In Raffaele Giancarlo and David Sankoff, editeurs, Combinatorial Pattern Matching, volume 1848 of *Lecture Notes in Computer Science*, pages 108–118. 2000. 70

[Gemert *et al.* 2008] Jan Van Gemert, Jan-Mark Geusebroek, Cor J. Veenman, Arnold W. M. Smeulders and Arnold W. M. Smeulders. *Kernel Codebooks for Scene Categorization.* In European Conference on Computer Vision (ECCV), pages 696–709, 2008. 35, 67

[Geusebroek *et al.* 2001] Jan-Mark Geusebroek, Rein van den Boomgaard, Arnold W.M. Smeulders and Hugo Geerts. *Color Invariance.* IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 12, pages 1338–1350, December 2001. 26

[Gotlieb & Kreyszig 1990] C. C. Gotlieb and Herbert E. Kreyszig. *Texture descriptors based on co-occurrence matrices.* Computer Vision, Graphics, and Image Processing, vol. 51, no. 1, pages 70–86, 1990. 18

[Guillaumin *et al.* 2010] Matthieu Guillaumin, Jakob Verbeek and Cordelia Schmid. *Multimodal semi-supervised learning for image classification.* June 2010. 4, 13, 64, 95

[Guo *et al.* 2010] Zhenhua Guo, Lei Zhang 0006, David Zhang and Xuanqin Mou. *Hierarchical multiscale LBP for face and palmprint recognition.* In ICIP, pages 4521–4524. IEEE, 2010. 20, 82

[Hamming 1950] Richard W. Hamming. *Error detecting and error correcting codes.* Bell System Technical Journal, 1950. 69

[Harris & Stephens 1988] Chris Harris and Mike Stephens. *A combined corner and edge detector.* In In Proc. of Fourth Alvey Vision Conference, pages 147–151, 1988. 23

[Hirst & St-Onge 1998] Graeme Hirst and David St-Onge. *Lexical chains as representations of context for the detection and correction of malapropisms.* In Christiane Fellbaum, editeur, WordNet: An Electronic Lexical Database, pages 305–332. MIT Press, 1998. 42

[Hofmann 1999a] Thomas Hofmann. *Probabilistic Latent Semantic Analysis.* In Proceedings of the Fifteenth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-99), pages 289–296, San Francisco, CA, 1999. Morgan Kaufmann. 6, 44, 45

[Hofmann 1999b] Thomas Hofmann. *Probabilistic latent semantic indexing.* In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '99, pages 50–57, New York, NY, USA, 1999. ACM. 6

[Huang *et al.* 1997] Jing Huang, S. Ravi Kumar, Mandar Mitra, Wei-Jing Zhu and Ramin Zabih. *Image Indexing Using Color Correlograms.* In Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97), CVPR '97, pages 762–. IEEE Computer Society, 1997. 17

[J. D. H. Farquhar & Shawe-Taylor 2005] Hongying Meng J. D. H. Farquhar Sandor Szedmak and John Shawe-Taylor. *improving "bag-of-keypoints" image categorisation generative models and pdf-kernels.* In Technical Report, University of Southampton, pages 40–42. Springer, 2005. 33, 37

[Jacob Goldberger & Greenspan 2003] Shiri Gordon Jacob Goldberger and Hayit Greenspan. *An efficient image similarity measure based on approximations of KL-divergence between two gaussian mixtures.* In In Proc. ICCV, pages 487–493, 2003. 37

[Jebara & Kondor 2003] Tony Jebara and Risi Kondor. *Bhattacharyya and Expected Likelihood Kernels.* In In Conference on Learning Theory. press, 2003. 37

[Jebara Tony & Andrew 2004] Kondor Risi Jebara Tony and Howard Andrew. *Probability Product Kernels.* J. Mach. Learn. Res., vol. 5, pages 819–844, December 2004. 38

[Jones 1972] Karen Sp?rck Jones. *A statistical interpretation of term specificity and its application in retrieval.* Journal of Documentation, vol. 28, no. 1, 1972. 40

[Jurie & Triggs 2005] Frederic Jurie and Bill Triggs. *Creating Efficient Codebooks for Visual Recognition.* In Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1 - Volume 01, ICCV '05, pages 604–610, Washington, DC, USA, 2005. IEEE Computer Society. 32

[Kumar & Annie 2012] D. Ashok Kumar and M.C. Loraine Charlet Annie. *CLUSTERING DICHOTOMOUS DATA FOR HEALTH CARE.* In International Journal of Information Sciences and Techniques (IJIST), pages Vol.2, No.2, 2012. 69

[Kurani *et al.* 2004] Arati S. Kurani, Dong hui Xu, Jacob Furst and Daniela Stan Raicu. *Raicu . Co-occurrence Matrices for Volumetric Data. The.* In 7th IASTED International Conference on Computer Graphics and Imaging, Kauai, 2004. 18

[Lanckriet *et al.* 2004] Gert R. G. Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui and Michael I. Jordan. *Learning the Kernel Matrix with Semidefinite Programming.* J. Mach. Learn. Res., vol. 5, pages 27–72, December 2004. 58

[Lategahn *et al.* 2010] Henning Lategahn, Sebastian Groß, Thomas Stehle and Til Aach. *Texture Classification by Modeling Joint Distributions of Local Patterns with Gaussian Mixtures.* IEEE Transactions on Image Processing, vol. 19, no. 6, pages 1548–1557, 2010. 87

[Lazebnik *et al.* 2006] Svetlana Lazebnik, Cordelia Schmid and Jean Ponce. *Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories.* In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2, CVPR '06, pages 2169–2178, Washington, DC, USA, 2006. IEEE Computer Society. 35, 36

[Leacock & Chodorow 1998] C. Leacock and M. Chodorow. *Combining local context and WordNet similarity for word sense identification.* In Christiane Fellfaum, editeur, MIT Press, pages 265–283, Cambridge, Massachusetts, 1998. 42

[Liao *et al.* 2007] ShengCai Liao, XiangXin Zhu, Zhen Lei, Lun Zhang and Stan Z. Li. *Learning Multi-scale Block Local Binary Patterns for Face Recognition.* In Seong-Whan Lee and Stan Z. Li, editeurs, ICB, volume 4642 of *Lecture Notes in Computer Science*, pages 828–837. Springer, 2007. 8, 20, 82

[Lin *et al.* 2007] Yen-Yu Lin, Tyng-Luh Liu and Chiou-Shann Fuh. *Local Ensemble Kernel Learning for Object Category Recognition.* 2012 IEEE Conference on Computer Vision and Pattern Recognition, vol. 0, pages 1–8, 2007. 10, 14

[Lin 1998] Dekang Lin. *An Information-Theoretic Definition of Similarity.* In In Proceedings of the 15th International Conference on Machine Learning, pages 296–304. Morgan Kaufmann, 1998. 42

[Lindeberg 1998] Tony Lindeberg. *Feature Detection with Automatic Scale Selection.* International Journal of Computer Vision, vol. 30, pages 79–116, 1998. 23

[Liu & Perronnin 2008] Yan Liu and Florent Perronnin. *A similarity measure between unordered vector sets with application to image categorization.* In IEEE Conference on Computer Vision and Pattern Recognition, 2008. 37

[Liu *et al.* 2009] Jingen Liu, Yang Yang and Mubarak Shah. *Learning semantic visual vocabularies using diffusion distance.* In CVPR, pages 461–468. IEEE, 2009. 32

[Liu *et al.* 2011a] Ningning Liu, Emmanuel Dellandrea, Bruno Tellez and Liming Chen. *Associating textual features with visual ones to improve affective image classification.* In International Conference on Affective Computing and Intelligent Interaction (ACII2011), pages 195–204, October 2011. 96

[Liu *et al.* 2011b] Ningning Liu, Yu Zhang, Emmanuel Dellandréa, Stéphane Bres and Liming Chen. *LIRIS-Imagine at ImageCLEF 2011 Photo Annotation Task.* In Vivien Petras, Pamela Forner and Paul D. Clough, editeurs, CLEF (Notebook Papers/Labs/Workshop), 2011. 102, 105, 118

[Liu *et al.* 2013] Ningning Liu, Emmanuel Dellandrea, Liming Chen, Chao Zhu, Yu Zhang, Charles-Edmond Bichot, Stephane Bres and Bruno Tellez. *Multimodal Recognition of Visual Concepts using Histograms of Textual Concepts and Selective Weighted Late Fusion Scheme.* Computer Vision and Image Understanding, vol. 117, no. 5, pages 493–512, May 2013. 42, 58

[Lloyd 1982] Stuart P. Lloyd. *Least squares quantization in pcm.* IEEE Transactions on Information Theory, vol. 28, pages 129–137, 1982. 31

[Lowe 2004a] David G. Lowe. *Distinctive Image Features from Scale-Invariant Keypoints.* Int. J. Comput. Vision, vol. 60, no. 2, pages 91–110, November 2004. 5, 25, 66, 89

[Lowe 2004b] David G. Lowe. *Distinctive Image Features from Scale-Invariant Keypoints.* Int. J. Comput. Vision, vol. 60, no. 2, pages 91–110, November 2004. 14

[MacQueen 1967] J. B. MacQueen. *Some Methods for Classification and Analysis of MultiVariate Observations.* In L. M. Le Cam and J. Neyman, editeurs, Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability, volume 1, pages 281–297. University of California Press, 1967. 31

[Maree *et al.* 2005] Raphael Maree, Pierre Geurts, Justus H. Piater and Louis Wehenkel. *Random Subwindows for Robust Image Classification.* In CVPR (1), pages 34–40. IEEE Computer Society, 2005. 24

[Martinet *et al.* 2011] Jean Martinet, Yves Chiaramella and Philippe Mulhem. *A relational vector space model using an advanced weighting scheme for image retrieval.* Information Processing and Management, vol. 47, no. 3, pages 391–414, 2011. 14

[McCallum & Nigam 1998] Andrew McCallum and Kamal Nigam. *A Comparison of Event Models for Naive Bayes Text Classification.* In Learning for Text Categorization: Papers from the 1998 AAAI Workshop, pages 41–48, 1998. 29

[Mikolajczyk & Schmid 2001] Krystian Mikolajczyk and Cordelia Schmid. *Indexing Based on Scale Invariant Interest Points.* In ICCV, pages 525–531, 2001. 24, 113

[Mikolajczyk & Schmid 2004] Krystian Mikolajczyk and Cordelia Schmid. *Scale and affine invariant interest point detectors.* International Journal of Computer Vision, vol. 60, no. 1, pages 63–86, 2004. 24

[Mikolajczyk & Schmid 2005] Krystian Mikolajczyk and Cordelia Schmid. *A performance evaluation of local descriptors.* IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 27, no. 10, pages 1615–1630, 2005. 28

[Miller *et al.* 1990] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross and Katherine Miller. *WordNet: An on-line lexical database.* International Journal of Lexicography, vol. 3, pages 235–244, 1990. 41

[Mori 2002] Tatsunori Mori. *Information Gain Ratio As Term Weight: The Case of Summarization of IR Results.* In Proceedings of the 19th International Conference on Computational Linguistics - Volume 1, COLING '02, pages 1–7, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. 44

[Nagel *et al.* 2011] Karolin Nagel, Stefanie Nowak, Uwe Kühhirt and Kay Wolter. *The Fraunhofer IDMT at ImageCLEF 2011 Photo Annotation Task.* In Vivien Petras, Pamela Forner and Paul D. Clough, editeurs, CLEF (Notebook Papers/Labs/Workshop), 2011. 102, 105, 118

[Nedeljkovic 2004] I. Nedeljkovic. *Image classification based on fuzzy logic.* In in Proc. Geo-Imagery Bridging Continents XXth ISPRS Congress, pages 12–23, 2004. 128

[Ojala *et al.* 2002a] Timo Ojala, Matti Pietikäinen and Topi Mäenpää. *Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns.* IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 7, pages 971–987, July 2002. 7, 19, 20, 81, 82

[Ojala *et al.* 2002b] Timo Ojala, Matti Pietikäinen and Topi Mäenpää. *Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns.* IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 7, pages 971–987, July 2002. 14, 66

[Ozuysal *et al.* 2010] Mustafa Ozuysal, Michael Calonder, Vincent Lepetit and Pascal Fua. *Fast Keypoint Recognition Using Random Ferns.* IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 3, pages 448–461, March 2010. 81, 83, 86

[Parikh & Zitnick 2010] Devi Parikh and C. Lawrence Zitnick. *The role of features, algorithms and data in visual recognition.* In CVPR, pages 2328–2335. IEEE, 2010. 64

[Paulhac *et al.* 2008] Ludovic Paulhac, Pascal Makris and Jean-Yves Ramel. *Comparison Between 2D and 3D Local Binary Pattern Methods for Characterisation of Three-Dimensional Textures.* In Proceedings of the 5th International Conference on Image Analysis and Recognition, ICIAR '08, pages 670–679, Berlin, Heidelberg, 2008. Springer-Verlag. 20

[Paulhac *et al.* 2009] Ludovic Paulhac, Pascal Makris and Jean-Yves Ramel. *A Solid Texture Database for Segmentation and Classification Experiments.* In Proceedings of the Fourth International Conference, pages 135–141, 2009. 20

[Pedersen *et al.* 2004] Ted Pedersen, Siddharth Patwardhan and Jason Michelizzi. *WordNet::Similarity: Measuring the Relatedness of Concepts.* In Demonstration Papers at HLT-NAACL 2004, HLT-NAACL–Demonstrations '04, pages 38–41, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics. 42

[Perronnin *et al.* 2006] Florent Perronnin, Christopher Dance, Gabriela Csurka and Marco Bressan. *Adapted vocabularies for generic visual categorization.* In In ECCV, pages 464–475, 2006. 32, 33

[Perronnin *et al.* 2010] Florent Perronnin, Jorge Sánchez and Thomas Mensink. *Improving the Fisher Kernel for Large-scale Image Classification.* In Proceedings of the 11th European Conference on Computer Vision: Part IV, ECCV'10, pages 143–156, Berlin, Heidelberg, 2010. Springer-Verlag. 5, 29, 35, 67

[Philbin *et al.* 2008] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, Andrew Zisserman and Andrew Zisserman. *Lost in quantization: Improving particular object retrieval in large scale image databases.* In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008. 67

[Rubenstein & Goodenough 1965] Herbert Rubenstein and John B. Goodenough. *Contextual Correlates of Synonymy.* Commun. ACM, vol. 8, no. 10, pages 627–633, October 1965. 42

[Salton & Buckley 1988] Gerard Salton and Christopher Buckley. *Term-weighting approaches in automatic text retrieval.* Information Processing and Management: an International Journal, vol. 24, pages 513–523, 1988. 40

[Salton & Mcgill 1986] G. Salton and M. J. Mcgill. Introduction to modern information retrieval. McGraw-Hill, Inc., New York, NY, USA, 1986. 39

[Salton *et al.* 1975] G. Salton, A. Wong and C. S. Yang. *A vector space model for automatic indexing.* Commun. ACM, vol. 18, no. 11, pages 613–620, November 1975. 6

[Sang & Meulder 2003] Erik F. Tjong Kim Sang and Fien De Meulder. *Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition.* In Proceedings of CoNLL-2003, pages 142–147, 2003. 131

[Schmidhuber 2012] Jurgen Schmidhuber. *Multi-column Deep Neural Networks for Image Classification.* In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), CVPR '12, pages 3642–3649. IEEE Computer Society, 2012. 128

[Schwab *et al.* 2002] Didier Schwab, Mathieu Lafourcade and Violaine Prince. *Antonymy and Conceptual Vectors.* In Proceedings of the 19th International Conference on Computational Linguistics - Volume 1, COLING '02, pages 1–7, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. 43

[Shan *et al.* 2009] Caifeng Shan, Shaogang Gong and Peter W. McOwan. *Facial expression recognition based on Local Binary Patterns: A comprehensive study.* Image Vision Comput., vol. 27, no. 6, pages 803–816, May 2009. 20, 82

[Shraddha Pandit 2011] Suchita Gupta Shraddha Pandit. *A COMPARATIVE STUDY ON DISTANCE MEASURING APPROACHES FOR CLUSTERING.* In International Journal of Research in Computer Science, volume 2, pages 29–31. White Globe Publications, 2011. 71

[Siddiquie *et al.* 2009] Behjat Siddiquie, Shiv Naga Prasad Vitaladevuni and Larry S. Davis. *Combining multiple kernels for efficient image classification.* In WACV, pages 1–8. IEEE Computer Society, 2009. 10

[Sivic & Zisserman 2003] J. Sivic and A. Zisserman. *Video Google: A Text Retrieval Approach to Object Matching in Videos.* In Proceedings of the International Conference on Computer Vision, volume 2, pages 1470–1477, 2003. 29

[Stricker & Orengo 1995] M. Stricker and M. Orengo. *Similarity of Color Images.* In Proc. SPIE Storage and Retrieval for Image and Video Databases, 1995. 16

[Su & Jurie 2011] Yu Su and Frédéric Jurie. *Semantic Contexts and Fisher Vectors for the ImageCLEF 2011 Photo Annotation Task.* In Vivien Petras, Pamela Forner and Paul D. Clough, editeurs, CLEF (Notebook Papers/Labs/Workshop), 2011. 120

[Swain & Ballard 1991] Michael J. Swain and Dana H. Ballard. *Color indexing.* Int. J. Comput. Vision, vol. 7, no. 1, pages 11–32, November 1991. 16, 21

[Tiberius Strat *et al.* 2012] Sabin Tiberius Strat, Alexandre Benoit, Herve Bredin, Georges Quenot and Patrick Lambert. *Hierarchical late fusion for concept detection in videos.* In ECCV 2012, Workshop on Information Fusion in Computer Vision for Concept Recognition, Firenze, Italy, oct 2012. 58

[Tola *et al.* 2010] E. Tola, V. Lepetit and P. Fua. *DAISY: An Efficient Dense Descriptor Applied to Wide Baseline Stereo.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 5, pages 815–830, May 2010. 27

[Tuceryan & Jain 1998] Mihran Tuceryan and Anil K. Jain. *Texture Analysis.* In Handbook of Pattern Recognition and Computer Vision, pages 235–276, 1998. 81

[Tuytelaars & Mikolajczyk 2007] Tinne Tuytelaars and Krystian Mikolajczyk. *Local Invariant Feature Detectors: A Survey.* Foundations and Trends in Computer Graphics and Vision, vol. 3, no. 3, pages 177–280, 2007. 23, 24

[van de Sande & Snoek 2011] Koen E. A. van de Sande and Cees G. M. Snoek. *The University of Amsterdam's Concept Detection System at ImageCLEF 2011.* In Vivien Petras, Pamela Forner and Paul D. Clough, editeurs, CLEF (Notebook Papers/Labs/Workshop), 2011. 120

[van de Sande *et al.* 2010] K. E. A. van de Sande, T. Gevers and C. G. M. Snoek. *Evaluating Color Descriptors for Object and Scene Recognition.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 9, pages 1582–1596, 2010. 23, 25, 67, 85, 113

[van Gemert *et al.* 2008] Jan C. van Gemert, Jan-Mark Geusebroek, Cor J. Veenman and Arnold W. M. Smeulders. *Kernel codebooks for scene categorization.* In ECCV 2008, PART III. LNCS, pages 696–709. Springer, 2008. 5, 29, 33, 34

[van Gemert *et al.* 2010] Jan van Gemert, Cor J. Veenman, Arnold W. M. Smeulders and Jan-Mark Geusebroek. *Visual Word Ambiguity.* IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, pages 1271–1283, 2010. 35

[Vasconcelos *et al.* 2004] Nuno Vasconcelos, Purdy Ho and Pedro J Moreno. *The Kullback-Leibler Kernel as a Framework for Discriminant and Localized Representations for Visual Recognition.* In Tomas Pajdla and Jiri Matas, editeurs, ECCV, volume 3023 of *Lecture Notes in Computer Science*, pages 430–441. Springer, 2004. 37

[Vasconcelos 2004] Nuno Vasconcelos. *On the efficient evaluation of probabilistic similarity functions for image retrieval.* IEEE Transactions on Information Theory, vol. 50, no. 7, pages 1482–1496, 2004. 37

[Vimal *et al.* 2008] Ankita Vimal, Satyanarayana R Valluri and Kamalakar Karlapalem. *An Experiment with Distance Measures for Clustering*, 2008. 69, 71

[Vishwanathan *et al.* 2010] S. V. N. Vishwanathan, Z. Sun, N. Theera-Ampornpunt and M. Varma. *Multiple Kernel Learning and the SMO Algorithm.* In Advances in Neural Information Processing Systems, December 2010. 59, 115

[Wang *et al.* 2009] Gang Wang, Derek Hoiem and David Forsyth. *Building text features for object image classification.* In CVPR, pages 1367–1374. IEEE, 2009. 4, 64

[Wang *et al.* 2010] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang and Yihong Gong. *Locality-constrained linear coding for image classification.* In IN: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN CLASSIFICATOIN, 2010. 5, 29, 35

[Winn *et al.* 2005] J. Winn, A. Criminisi and T. Minka. *Object categorization by learned universal visual dictionary.* In In ICCV, pages 1800–1807. IEEE Computer Society, 2005. 33

[Xioufis *et al.* 2011] Eleftherios Spyromitros Xioufis, Konstantinos Sechidis, Grigorios Tsoumakas and Ioannis P. Vlahavas. *MLKD's Participation at the CLEF 2011 Photo Annotation and Concept-Based Retrieval Tasks.* In Vivien Petras, Pamela Forner and Paul D. Clough, editeurs, CLEF (Notebook Papers/Labs/Workshop), 2011. 102, 105, 118

[Yang & Pedersen 1997] Yiming Yang and Jan O. Pedersen. *A Comparative Study on Feature Selection in Text Categorization.* In Proceedings of the Fourteenth International Conference on Machine Learning, ICML '97, pages 412–420, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc. 44

[Yang *et al.* 2007] Jun Yang, Yu-Gang Jiang, Alexander G. Hauptmann and Chong-Wah Ngo. *Evaluating Bag-of-visual-words Representations in Scene Classification.* In Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval, MIR '07, pages 197–206, New York, NY, USA, 2007. ACM. 30

[Yilmaz *et al.* 2008] Emine Yilmaz, Evangelos Kanoulas and Javed A. Aslam. *A Simple and Efficient Sampling Method for Estimating AP and NDCG.* In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '08, pages 603–610, New York, NY, USA, 2008. ACM. 32

[Yue *et al.* 2007] Yisong Yue, Thomas Finley, Filip Radlinski and Thorsten Joachims. *A support vector method for optimizing average precision.* In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '07, pages 271–278, New York, NY, USA, 2007. ACM. 74, 89, 101, 117

[Zhang *et al.* 2000] Dengsheng Zhang, Aylwin Wong, Maria Indrawan and Guojun Lu. *Content-based Image Retrieval Using Gabor Texture Features.* In IEEE Transactions PAMI, pages 13–15, 2000. 81

[Zhang *et al.* 2007] Jianguo Zhang, Marcin Marszaek, Svetlana Lazebnik and Cordelia Schmid. *Local features and kernels for classification of texture and object categories: a comprehensive study.* volume 73, pages 213–238, 2007. 32

[Zheng 2004] Zhaohui Zheng. *Feature selection for text categorization on imbalanced data.* ACM SIGKDD Explorations Newsletter, vol. 6, page 2004, 2004. 44

[Zhou *et al.* 2010] Xi Zhou, Kai Yu, Tong Zhang, Thomas S. Huang and Thomas S. Huang. *Image Classification Using Super-Vector Coding of Local Image Descriptors.* In European Conference on Computer Vision, pages 141–154, 2010. 5, 29, 35

[Zhu *et al.* 2010] Chao Zhu, Charles-Edmond Bichot and Liming Chen. *Multi-scale Color Local Binary Patterns for Visual Object Classes Recognition.* In IEEE, editeur, International Conference on Pattern Recognition (ICPR), pages 3065–3068, August 2010. 8, 20, 81, 82, 113

[Zhu *et al.* 2011] Chao Zhu, Charles-Edmond Bichot and Liming Chen. *Visual object recognition using daisy descriptor.* In IEEE, editeur, IEEE International Conference on Multimedia and Expo (ICME), July 2011. 5, 67, 76