1-1-2011

Combining visual features and contextual information for image retrieval and annotation

Rui Zhang
Ryerson University

Combining Visual Features and Contextual Information for Image Retrieval and Annotation

by

Rui Zhang

Master of Engineering, Tianjin University, 2004

Bachelor of Engineering, Tianjin University, 2002

A dissertation

presented to Ryerson University

in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in the Program of

Electrical and Computer Engineering

Toronto, Ontario, Canada, 2011

I hereby declare that I am the sole author of this dissertation.

I authorize Ryerson University to lend this dissertation to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this dissertation by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Combining Visual Features and Contextual Information for Image Retrieval and Annotation

Doctor of Philosophy 2011

Rui Zhang

Electrical and Computer Engineering

Ryerson University

Abstract

This thesis is primarily focused on the information combination at different levels of a statistical pattern classification framework for image annotation and retrieval. Based on the previous study within the fields of image annotation and retrieval, it has been well-recognized that the low-level visual features, such as color and texture, and high-level features, such as textual description and context, are distinct yet complementary in terms of their distributions and the corresponding discriminative powers for dealing with machine-based recognition and retrieval tasks. Therefore, effective feature combination for image annotation and retrieval has become a desirable and promising perspective from which the semantic gap can be further bridged. Motivated by this fact, the combination of the visual and context modalities and that of different features in the visual domain are tackled by developing two statistical pattern classification approaches considering that the features of the visual modality and those across different modalities exhibit different degrees of heterogeneities, and thus, should be treated differently. Regarding the cross-modality feature combination, a Bayesian framework is proposed to integrate visual content and context, which has been applied to various image annotation and retrieval

frameworks. In terms of the combination of different low-level features in the visual domain, the problem is tackled with a novel method that combines texture and color features via a mixture model of their joint distribution. To evaluate the proposed frameworks, many different datasets are employed in the experiments, including the COREL database for image retrieval and the MSRC, LabelMe, PASCAL VOC2009, and an animal image database collected by ourselves for image annotation. Using various evaluation criteria, the first framework is shown to be more effective than the methods purely based on the low-level features or high-level context. As for the second, the experimental results demonstrate not only its superior performance to other feature combination methods but also its ability to discover visual clusters using texture and color simultaneously. Moreover, a demo search engine based on the Bayesian framework is implemented and available online.

Acknowledgements

First and foremost, I would like to express my sincere acknowledgement in the advice and help of my Ph.D. supervisor, Dr. Ling Guan. Professor Guan has supported me throughout my Ph.D. study and thesis-writing period with his patience and knowledge whilst allowing me the room to work in my own way. I appreciate all his contributions of time, ideas, and funding to make my Ph.D. experience stimulating and productive. His enthusiasm about his research was extremely contagious and motivational for me.

I also would like to thank the many people with whom I have collaborated in the research during the past few years. Particularly, I am indebted to my mentor Dr. Lei Zhang and my colleague Dr. Xin-Jing Wang at Microsoft Research Asia (MSRA), who shared with me a great deal of expert knowledge and experience in the field of visual recognition and image understanding. They have contributed immensely to my personal and professional time during my internship at MSRA. In addition, it is my pleasure to thank Professor Kim-Hui Yap and Dr. Kui Wu at Nanyang Technological University for the enormous insightful and detailed discussion on the research topic of image annotation.

Being a member of the Ryerson Multimedia Research Laboratory (RML) has been and will always be my spiritual wealth because it is such a great source of friendships as well as good advice and collaboration. In my daily work I have been blessed with a friendly and cheerful group of fellow students.

For this dissertation, my gratitude also goes to the defence committee members: Professor Jimmy Huang, Professor Chen (Cherie) Ding, Professor Sri Krishnan, Professor Javad Alirezaie, Professor Lian Zhao, Professor Olivia Das and Dr. Jennifer Mactavish.

Last but not least, I would like to thank my parents for all their love and encouragement and for the support in all my pursuits.

Contents

List of Tables

List of Figures

List of Appendices

Chapter 1

Introduction

1.1 Background

The ever-lasting growth of multimedia information has been witnessed and experienced by human beings since the beginning of the information age. An immediate challenge resulting from the information explosion is how to intelligently manage and enjoy the multimedia databases. Among the technologies developed in response to this challenge, multimedia information retrieval (MIR) has emerged as a critical tool for the access to the multimedia content relevant to a user's information need from a large scale source of information, which is commonly considered to be the Internet nowadays. The applications of MIR can be related to many different aspects of our life, such as searching for travel information, quick access of library catalog and other educational resources, and building online social network, just to name a few.

By far, most of the commercial search engines, such as Google™, Microsoft Bing™, Yahoo™, and Baidu™, rely on textual information to index and search for the available multimedia content. Being completely indexed and searched with textual information

has several limitations.

- First, unless manually annotated by professionals or very conscientious amateur multimedia producers, the textual information is usually very unreliable and unstructured, which may cause a great number of errors during the subsequent indexing and search processes. Even if such annotation is securable, different people still can use different words and phrases to describe the same semantic content. Besides, the description from one person may differ from that of another because of the subject's knowledge background, interest and attention, etc. Although machine translation can be employed to convert the annotation in one language to that in another, many techniques of this kind might not work well because the annotation of multimedia content is normally composed of isolated keywords or phrases rather than complete sentences.

- Second, there are three types of queries summarized in [1], which are browsing, category search and target search. The first type of query starts with no specific information need, i.e. a user just wants to browse the database to see if there is anything interesting to look into with more details. The second type of query aims at finding the objects or scene belonging to the same semantic category, e.g. finding all images of cars. The last kind of query requires the system to find images of the same object or exactly a copy of the query. Other types of queries can also be found in [2, 3]. Clearly, not all kinds of queries can be efficiently formulated using keywords or a phrase. For example, target search and category search can be best solved by means of directly expressing the information need using images and using keywords, respectively. In addition, if visual content analysis can embed some semantic structure into the unstructured image data, browsing can be made

much more efficient.

1.2 Challenges and Relevant Technologies

To compensate for the afore-mentioned drawbacks of the text-based search engines, there has been much research effort devoted to the development of search technologies without utilizing text as well as automatically assigning relevant textual description to images. In the course of the technological development of MIR, various approaches have been proposed with the ultimate goal of enabling semantic-based search and browsing.

1.2.1 Content-Based Image Retrieval

Among those intensively explored topics, content-based image retrieval (CBIR), born at the crossroad of computer vision, machine learning and database technologies, has been studied for more than a decade, yet still remaining difficult [1, 4]. In a nutshell, the content-based approaches to image retrieval primarily rely on the pictorial information, a.k.a. low level visual features such as color, texture, shape and layout, which can be automatically extracted from images for similarity measure. Two well-recognized extremely challenging issues associated with CBIR are the sensory gap and semantic gap, as illustrated in Fig. 1.1. The sensory gap refers to the difference between the appearance of an object in a real scene and its numerical representation extracted from the information captured with sensors. Information can be lost during both the capturing process and the feature extraction stage. The semantic gap means the disparity between the numerical representation extracted from the recorded information and the interpretation of the recorded data. In terms of the semantic gap, low-level visual features accurately characterizing the semantic meaning of images are difficult to discover, which is coupled

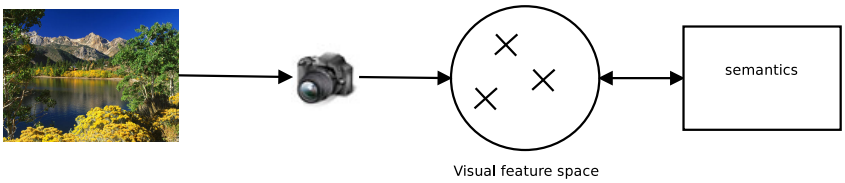


Figure 1.1: The illustration of the sensory and semantic gap.

with the imperfection of the conventional distance functions embedded in the Euclidean space. These two problems lead to the observation that semantically relevant images may be located far away from each other in the space of the low-level visual features. According to [1], the sensory and semantic gaps are very wide when the problem domain encompasses a large number of object categories, highly variable illumination conditions and viewing angles. As a result, various photometric and geometric changes lead to quite complex distribution of the numerical representation of image data, typically in a very high dimensional space.

From the perspective of machine learning, if images are represented using numerical descriptors in a feature space where a distance metric is defined, searching for similar images inevitably boils down to the problem of measuring the distance between images or how well a candidate image is aligned with the model characterizing the information need represented by a query. Early retrieval paradigm for content-based search engine design is query-by-example, i.e. a user initializes a query session by providing the system with an image containing the semantic meaning relevant to the his/her information need. The selected distance function measures the similarity between the example image and each of candidate images in the database. The output of the distance function is used to rank the candidate images, of which the most similar (in the sense of being closest to the query in the feature space) ones are returned to the user. This rudimentary procedure describes the minimal workload for retrieving images based on the visual content similarity and

generally can not satisfy the users unless the image domain is narrow and the distance function is designed properly based on the domain knowledge. To be specific, the user's query may be very complex and hence difficult to be represented using a single image. More often than not, the visual content of an image is either insufficient or excessive for expressing the semantic meaning of the actual information need. As an example of having too much irrelevant visual content, if an image of a car in a cluttered scene is used as the query image for a category search, the visual information of the objects other than the car in the image will make negative contribution to the query formulation. As a result, it reduces the discriminative power of the image representation and the similarity measure function employed in a search engine. Moreover, it is worth noting that CBIR, as a pattern classification problem, is rather different from others in that a semantic class of interest to a user (what to search for) during the system operation time is not pre-defined and remains unknown until a user submits queries for them.

To deal with this issue, an online learning approach called relevance feedback (RF) was borrowed from the domain of document retrieval. Through human machine interaction (HMI), candidate images labeled by the users during the RF process can be used to refine the query formulation, i.e. recalculating the feature representation of the query or the parameters of the model distinguishing the relevant images from irrelevant ones. Although RF is capable of alleviating the semantic gap to some extent, the amount of time needed for learning the semantic meaning in a previously unseen query/semantic class online is fairly limited because users normally expect an efficient search engine requiring little HMI. This is the most difficult problem associated with the RF techniques in that very few labeled data can be obtained for subsequent online learning within a short period.

To tackle the challenges which still remain after incorporating RF techniques, we

propose a framework which integrates learning from the information accumulated through a long-term period with the conventional CBIR approaches. This leads to an image retrieval framework utilizing both low-level visual content and high-level context, with the latter estimated using the past search results.

1.2.2 Automatic Image Annotation

Another promising key technology for more effective search engine design is the construction of a mapping from visual content to high-level semantics, i.e. automatic image annotation. The significance of this technology lies in its usefulness of enriching the multimedia databases labeled with textual descriptions. This will considerably improve the efficiency and effectiveness of the development and deployment of the state-of-the-art keyword-based search technologies. The problems to be handled by this technology constitute a super set of the problems collectively known as object recognition. According to the research in the field of psychology and computer vision, scene perception can be rapidly performed without analyzing the details of individual objects. However, the detailed description on the semantic meaning still largely relies on detecting and recognizing the objects present in a scene. A number of generative machine learning frameworks have been developed to model the visual feature representation of images and leveraged to assign semantic labels to them. Such frameworks can formulate the parametric form of a model by simulating a conceptual data generation process, which results in much insight into the discovery of meaningful structure of complicated image data.

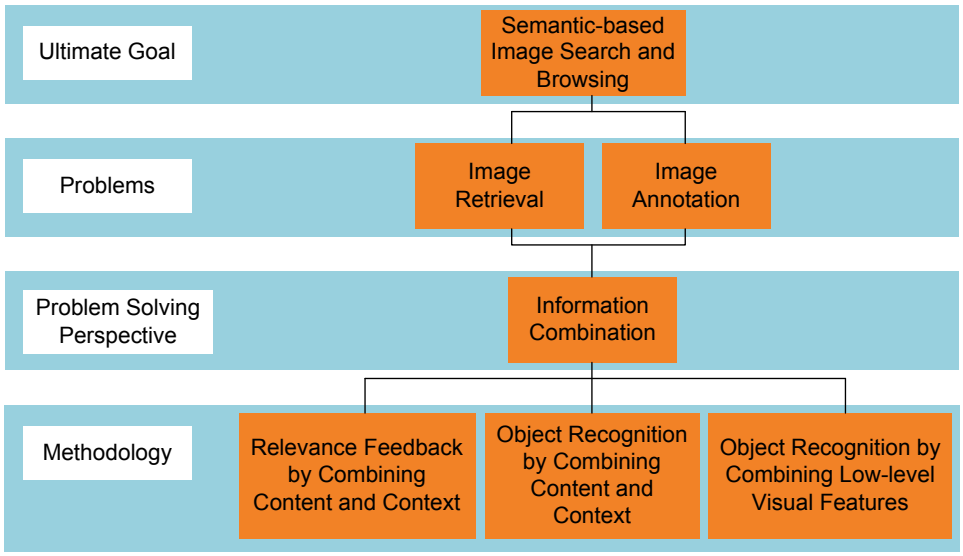


Figure 1.2: The general overview of the thesis.

1.3 Overview and Contributions of the Thesis

In accordance with the above-mentioned background and challenges, some exploratory research tackling the major issues with respect to the research topics related to MIR is presented in this thesis, with the ultimate goal of developing technologies to enable semantic-based image search and browsing. Illustrated in Fig. 1.2 is the general overview of the thesis. Throughout years of intensive study on visual recognition, it has been well-recognized within the research community that utilizing the information from multiple modalities and multiple types of features in each modality results in performance improvement [5]. Intuitively, as long as different modalities and descriptors contain complementary discriminative information, the more of them are combined into the classification process, the more semantic gap may be bridged. Nonetheless, how much improvement can be acquired still depends on how appropriately the various sources of information are jointly exploited. Based on this line of thought, we have taken the perspective of

combining multiple sources of information to solve the technical problems of image retrieval and annotation. In terms of the specific methodology, we have proposed statistical frameworks to combine 1) multiple modalities, i.e. visual content and context, and 2) different features belonging to the visual feature modality. The proposed approaches are intended to handle the feature combination at different levels of a pattern recognition process. To be specific, the combination across different modalities can be regarded as a high level fusion because the input of the fusion scheme consists of the output of the model of each individual modality. For this kind of task, a general Bayesian framework is developed, where the features of different modalities are modeled separately yet used jointly through the Bayes' theorem. On the contrary, because of the stronger correlation among different visual features compared with that of different modalities, the combination of various features in the visual modality is solved through low-level fusion, where a single statistical model is utilized for jointly modeling all the features. It is also noteworthy that the low-level fusion can be further integrated with the high-level one in a more general pattern classification system. In this case, the low-level fusion plays the role of content analysis in the overall framework.

The technical contribution of the thesis is summarized as follows.

- A general Bayesian framework is proposed. It integrates the content analysis for the likelihood evaluation and the context analysis for estimating the *a priori* probabilities. The latter is based on the maximum entropy estimation of the statistical dependence across multiple entities, either an image or a semantic class, which is defined as the context in the presented study.

- The Bayesian framework is applied to the image annotation problem, where the content analysis and context analysis are complementary to each other, result-

ing in the performance that is superior to both content-based and context-based approaches. Experimental results show when one component, either content or context, failed to assign a label to the image, the other component will help correct the mistake. By further incorporating a content-based search module to the Bayesian image annotation framework, users are not involved in the annotation process to provide feedback on the list of relevant semantic classes ranked using the output of the content-base component. Hence, the Bayesian image annotation framework is fully automatic.

- The Bayesian framework is also applied to the image retrieval problem. Considering the limited period for learning semantics from users' queries and RF steps, the statistical correlation across candidate images is learned through a long-term process, and used as the contextual information to boost the performance of the content-based search. Results demonstrate that the retrieval performance reach the same level of accuracy as that of purely content-based approaches at a considerably higher speed. This also implies the possibility of letting past users help prospective users using such a statistical framework, i.e. a collaborative search engine. A prototype image retrieval system has been implemented for subjective evaluation of the retrieval framework.

- Taking into consideration the fact that there are objects which can produce characteristic sound, the audio feature is utilized as another type of contextual information to solve the semantic gap problem of image retrieval. Here, the sound is considered as the holistic background context, as opposed to the statistical relation between different object categories or images. As an additional application of the Bayesian framework, it is employed to integrate the visual and audio modalities.

- Targeting more effective content analysis for automatic image annotation using multiple low-level visual features, a general statistical model, called multi-feature probabilistic latent semantic analysis (MF-pLSA), is proposed. It jointly characterizes the distributions of color and texture and decomposes them into a mixtures of topic distributions in the color and texture feature domains, respectively. This also leads to the joint clustering of the image data based on the distributions in the two feature spaces. An EM-based learning algorithm is derived for estimating the parameters of the MF-pLSA with a given training set. A maximum likelihood classification framework is designed to annotate objects in images using the MF-pLSA. Extensive experimental study on the MF-pLSA and the comparison with other methods of combining low-level visual features show that for most of the object categories under our consideration, the MF-pLSA outperforms others in terms of both recall and and precision.

1.4 Organization of the Thesis

The rest of the thesis is organized as follows. In Chapter 2, a detailed review on the existing works related to CBIR and automatic image annotation are discussed. In Chapter 3, we first introduce the general Bayesian framework for combining the content and context modalities. Then, its application to CBIR and image annotation are elaborated. Furthermore, by considering the audio modality as the background context, a Bayesian image retrieval framework using both visual and audio features is presented in Chapter 4. In Chapter 5, we move into a lower-level of feature combination, which aims at the combination of color and texture for automatic image annotation. Finally, the general conclusion and future research directions are provided in Chapter 6.

Chapter 2

Literature Review on Related Works

2.1 Introduction

In this chapter, a review on the existing approaches to the general domain image annotation and retrieval is presented. There is no doubt that the existing works regarding image retrieval and annotation are rather diversified in terms of their methodology because the subject matter of interest is born at the cross road of many research fields, such machine learning, computer vision and database technologies, just to name a few. Taking into account that the major theme of this thesis is the integration of multiple sources of information, i.e. (1) low level visual content and high level context and (2) different kinds of low level visual features, the primary criterion of the taxonomy of the methods covered in this chapter is the perspective through which they deal with the semantic gap. At the top level, existing works are divided into those explicitly solving image retrieval and those addressing image annotation. As to image retrieval, existing RF and multi-modal approaches are discussed. Regarding image annotation, previously developed methods are further categorized based on the consideration of context and the specific modeling

schemes.

2.2 Image Retrieval

Conventional CBIR systems exploiting low-level visual features, such as color, texture and shape, to represent images in a feature space. Early systems which successfully exploited these features include QBIC [6, 7], VisualSEEk [8], Photobook [9] and Virage [10]. These low-level feature-based systems have proven effective to the extent of pre-attentive similarity, which is largely due to the well-known sensory and semantic gaps. In recent years, these challenges have been addressed through the RF and multi-modal approaches.

2.2.1 Relevance Feedback

Noticing the critical role that human beings play in recognizing and comparing semantic content in multimedia objects, the human-supervised retrieval process was introduced, with the landmark being the application of RF to CBIR. The RF was originally developed for document retrieval. It aims at assisting users in formulating queries more accurately and comprehensively by integrating the relevant retrieved items with the initial query during the first few search iterations or retraining the model characterizing the information need iteratively. Various RF techniques proposed hitherto focused on three major aspects, i.e. the query point movement approaches [11], [12], the weighted distance methods [13, 14], and model-based similarity measure strategies. Regarding the last aspect, which is essentially the approximation of a function consistent with human visual perception, existing works can be divided into two categories. The first category is the nearest neighbor CBIR (NN-CBIR), in which the distance function is either lin-

ear [15, 16] or non-linear [17, 18]. In [17], an interactive neural network-based learning framework was developed, in which the most informative images are used to refine the query model. Based on a similar neural network approach, the method in [17] was generalized such that the system can accept multiple degrees of relevance in the RF stage, a.k.a a soft RF [18] scheme. The second category of approaches formulate image retrieval as a pattern classification problem, either a two-category classification [19] or a multi-category one [20]. In [19], the methodology of active learning was employed for acquiring training samples during RF. Specifically, instead of informing the system of the positive and negative images, users are asked to classify the most ambiguous one into either the relevant or the irrelevant category. The model of the query is then retrained with these samples that can not be confidently classified yet. We refer to the above RF techniques based on each individual user's feedback as short term relevance feedback (STRF) as the learning only continues until the end of a single search request. The approaches in this category alleviate the semantic gap by incorporating human users' knowledge into the process of labeling training samples yet still suffering from the problem of sample sparseness, as average users are normally willing to select only a few relevant and irrelevant images. In addition, as irrelevant images may be distinct from the relevant ones in many different ways, training samples of the two categories in the context of RF are very likely imbalanced. It should also be noted that the afore-mentioned RF techniques are complementary to each other rather than being alternative, in the sense that they can be integrated to achieve better performance, examples of which can be found in [12, 21].

Having recognized that learning the semantics of images is a long-term task, long term relevance feedback (LTRF) based on users' feedback across multiple queries has been proposed. In terms of the methods for extracting the information from retrieval history, singular value decomposition (SVD) was employed to construct a hidden semantic space

(HSS) [22], in which the retrieval history is recorded in a hidden semantic matrix and the classification is performed using support vector machine (SVM) in the dimensionality-reduced HSS. As another example, content-free image retrieval (CFIR) [23] directly exploiting statistical dependence across the images in a database was proposed, where semantically similar images identified by human users in the past retrieval sessions are connected to one another using a maximum entropy model. An inherent limitation of CFIR is the cold start problem, resulting from deficient or even unavailable training data. In [24], a new RF framework was proposed to facilitate continuously accumulating past retrieval results and in turn incrementally learning the high-level knowledge for CFIR.

2.2.2 Multi-modal Image Retrieval

Towards the goal of bridging the semantic gap, researchers have endeavored to explore multi-modal image retrieval employing both textual and visual information. For example, the difference between [25] and [26] is that they perform image-level and region-level keyword assignment, respectively. While [25] is essentially keyword-based retrieval with the content-based module playing a role of translating the query into keywords (online annotation), a linear late fusion was employed in [26] to combine the similarity measured in the textual and visual domains. It should be noted that the determination of the coefficients for the linear fusion is usually heuristic. As an image is inherently a 2-dimensional information carrier, little effort has been put on incorporating audio information into image retrieval when compared with video retrieval until recently [27,28]. It is because of the rocketing popularity of camera-integrated mobile phones and demands of users for on-the-go information retrieval that the audio-visual image retrieval has become a new research frontier. As an example, the fusion scheme employed in [27] is

similar to that in [26]. In addition, the system in [28] enables queries expressed using different individual modality.

2.3 Image Annotation

As mentioned earlier, query formulation based on the low-level visual content may be burdensome for human beings, especially when the images related to some high-level semantics are expected in a search result. Unlike low-level visual features, human language was created to record human knowledge and to express ourselves. Hence, textual information expressed using natural language lends itself to characterizing the semantics in images. Feng [29] proposed a multi-modal image retrieval, which enables both query-by-keyword and query-by-example with the measured similarity linearly combined at a subsequent step. Nonetheless, existing behind the performance improvement of the multi-modal retrieval is the lack of textual annotation with sufficient semantic richness. Considering the unreliability of the text surrounding the images on the Internet and the infeasibility of manually annotating large scale image databases, automatic image annotation intended to facilitate semantic-based search and browsing has been studied for years. Essentially, the technique underlying automatic image annotation is machine-based visual recognition, the goal of which is to automatically classify the input visual information into several predefined semantic categories. Based on the level of abstraction of the predefined semantic categories and the target of annotation, there are, in a coarse-to-fine order, three major tasks which can be considered as particular instances of automatic image annotation.

1. Scene Classification

This task involves the most abstract predefined categories, including locations and events. Early works, such as [30] and [31], use global features and hence only prove effective with respect to general semantic categories, such as indoor versus outdoor and city view versus natural scene. Advanced approaches developed recently, such as [32] and [33], are capable of handling the classification into more specific scene categories, e.g. inside of cities, streets, forest, coast, open country, bedroom, living room, etc. In addition, there are also approaches recognizing the events going on behind the visual scene when an image is taken. As an example, the method in [34] can distinguish between events such as badminton, rock climbing and snowboarding, etc.

2. Image-based Object Annotation

   Simply put, the goal of the second task is to answer if one or more instances of a certain type of objects exists in an image. The location and area of the objects are usually not considered. Methods falling into this category include, but not limited to, [35], [36], [37], [38], [39], [40]. While using global and/or local visual descriptors, these approaches associate words with an entire image by keeping the words on the top portion of a list ranked based on their *a posteriori* probabilities given the visual information.

3. Region-based Object Annotation

   Distinct from the above two tasks, region-based annotation with keywords corresponding to objects is the task aiming at generating the most specific correspondence between words and image structures. Examples of this category include [41], [42] and [43]. The latent variable models proposed in [44], [45] and [34]

are capable of dealing with the association of words with both images and their regions.

Having stated the tasks for automatic image annotation, an in-depth review on the existing approaches is provided in what follows. As a high-level machine vision problem, methods for image annotation can be classified based on many different aspects, including feature extraction, feature modeling, specific tasks to solve, etc. The methods reviewed here are organized in two parts, one of which covers the approaches which consider observed samples as independently and identically distributed (i.i.d.) data, and the other of which encompasses the methods taking into account the statistical dependency across observed samples. Within each part, the approaches are compared in terms of the other aspects of the methodology. Using the terminology of pattern recognition, the second category of approaches adopt the paradigm of context-aware classification. In accordance with our domain knowledge on visual recognition and as demonstrated in some research works, the second kind of approaches outperform those of the first kind by taking into account the cross-sample dependency. However, both of these two kinds of approaches have pros and cons, which can be better explained with Fig. 2.1. The directed graphical model on the left-hand side is a particular example of a hierarchical structure for modeling i.i.d samples. It can be used to capture the semantic structure of an image collection by modeling scene, object, part and visual descriptor using the nodes from top to bottom. In addition, the hierarchically structured model can be employed to solve all of the three afore-mentioned tasks of image annotation. The example of context-aware model shown on the right-hand side compensates for the disregard of the hierarchical model with respect to the cross-sample dependency. Nonetheless, this is achieved at the cost of higher computational complexity due to top-layer nodes' nature of being hidden

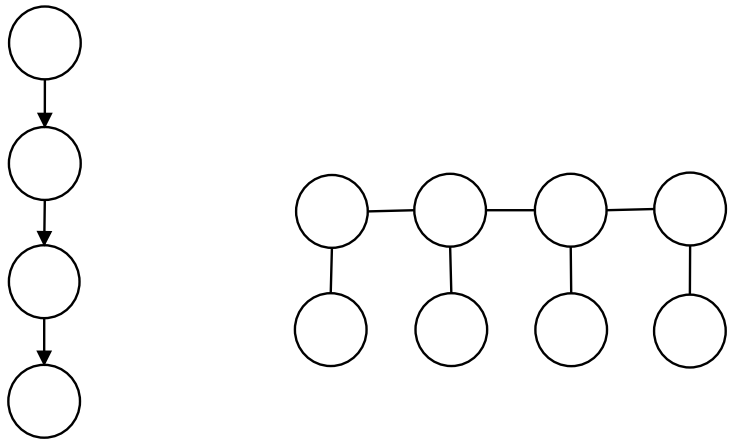


Figure 2.1: The comparison between an i.i.d. data model and a context-aware classification model. The i.i.d. data model shows the data generation process of a single sample of a generative model. On the right-hand side, the connection between the nodes on top layer of the context-aware model indicates their correlation.

variables and their marginalization during learning. Moreover, the optimal inference on the states of the top-layer nodes involves the enumeration over N^K possible cases, where N and K are the numbers of samples and category labels, respectively. Although there are efficient algorithms such as dynamic programming and more generally the sum-product algorithm [46] for factor graphs, they are only optimized for tree-structured graphs. It should be noted that the examples illustrated in Fig. 2.1 are only for comparing the merits and drawbacks of the two modeling strategy. In the literature, many variants of them have been proposed, with the major difference lying in the dependency among the variable nodes and the distribution they follow.

Prior to the discussion on previous methods, it is worth mentioning that the third task of automatic image annotation is close to object category recognition. Furthermore, there are also a great deal of common aspects in terms of the methodology dealing with these two subject matters. Therefore, previous approaches to object category recognition are also covered in order to give a comprehensive review upon the advancement of the

research field of image annotation.

2.3.1 Image Annotation by Modeling Independently and Identically Distributed Data

Existing approaches falling into this category can be further classified into supervised and unsupervised models. Using supervised models, some researchers define each keyword as a class, as in [38, 47], whereas others, as in [39, 48], consider a set of related keywords as a class, which is also referred to as a concept. For the latter, the keywords of the top ranked concept categories go through a keyword selection process and those selected are propagated to an image or a region to be annotated. With unsupervised models, approaches purely based on image data, e.g. [49, 50], do not pre-define a set of semantic classes. They rely on the algorithms to automatically discover the significant categories with coherent visual properties, which is essentially similar to a clustering process. Meanwhile, there has been a substantial amount of research works tackling image annotation through jointly modeling image data and text, e.g. [44, 45]. In this case, keywords are merely treated as observed samples from the domain of text, rather than labels corresponding to semantic classes. Therefore, they are referred to as multi-modal approaches to image annotation in the literature.

Unsupervised Approaches

An early unsupervised approach can be found in [51]. In this work, image patches obtained by grid-based uniform image partitioning are used as object regions. These patches are clustered through vector quantization. If a patch is assigned to a cluster, all of the keywords of the image, to which that patch belongs, are taken to that cluster as

well. This way, each cluster eventually has its own histogram of frequencies of occurrence of all the keywords. To annotate a new image, its patches are assigned to the clusters based on the nearest neighbor rule and the average of the histograms of the selected clusters are used to rank the keywords.

Motivated by the demand for such applications as efficient browsing of an image collection, query by text and automatic image annotation, a hierarchical model [52] was proposed to learn the semantic structure of a collection of images with associated textual labels. In this context, the semantic structure can be interpreted as the summary of the organization of an image collection based on their semantics. The ultimate goals of this work are: 1) to associate unlabeled images with textual annotation such that efficient query by text is enabled, and 2) to expose the semantic structure of an annotated image collection to users such that efficient browsing is supported. In light of the success of probabilistic latent semantic analysis (pLSA) in the research area of textual document[1] analysis, the hierarchical model in [52] was built based on the principle of pLSA. Each image segment or annotation keyword is generated by first selecting a cluster, denoted c, and then selecting a level of generality, denoted l, which together determine the distribution, denoted $P(x|c,l)$, of the segment or keyword, denoted x, conditional on the above selected quantities. In the area of machine learning, these conditional probabilities are referred to as aspects. In the area of document analysis, they are termed topics. Using the Expectation-Maximization (EM) procedure, the set of topics can be learned, which represent the semantic structure of an image collection. The probability of an image having its own keywords and segments, denoted $P(x|d)$, is represented by the product

[1]In the literature, many statistical models for textual document analysis have been borrowed to solve the analysis in image domain. Hence, document and image are usually used interchangeably. Instead of following this style of terminology, in the rest of the thesis, the term document is particularly used to refer to textual documents in order to distinguish them from images, unless stated otherwise.

of weighted sums of $P(x|c,l)$, taken into account the assumption of the i.i.d. nature of the samples. In addition, the weights quantify the decomposition of $P(x|d)$ with respect to the topics, which are considered as the signature of an image in the topic space. For the retrieval application, since each candidate image in the database has its own $P(x|d)$, the similarity between a query and a candidate image can be measured by calculating the probability that the set of keywords and segments, denoted X_q, of the query are generated by the $P(x|d)$ of candidate image d, i.e. $P(X_q|d)$. To annotated an image without any textual labels, the joint probability of a keyword w and all of the segments of the image, denoted S, can be calculated based on $P(x|c,l)$, which can be considered as the probability of matching a keyword with the set of segments of an image. The keywords with high matching probabilities are assigned to the image. In terms of low-level visual features, size, position, color, texture and shape descriptors are stored in a single vector. This kind of low-level feature combination is referred to as vector concatenation of individual descriptor vectors.

The above hierarchical model is a particular example of the unsupervised learning of the joint distribution of image and text using a pLSA-based topic model. The most fundamental principle of this kind of approach is that it learns the distribution of text and visual features of an image, i.e. $P(x|d)$, by decomposing it into a weighted sum of a set of topic distributions, represented as $P(x|O)$, where O denotes the set of variables collectively determining a unique topic. Meanwhile, the set of image-specific weights $P(O|d)$, or topic mixture coefficients, can be thought of as the signature of image d in the space spanned by the topic distributions. In the case of [52], $O = \{c, l\}$. Since the topic mixture coefficients of all images are parameters of the model in [52], a hierarchical model which drops the dependence of a topic mixture coefficient on an image was proposed in [53] to restrict the size of the model. In [45], a hierarchical model which includes image-specific

but cluster-independent topic mixture coefficients was introduced. It controls the size of the model yet retaining its ability to generate image signatures for retrieval applications. In [44], latent Dirichlet allocation (LDA) was employed to jointly modeling images and associated text. The difference between LDA and pLSA is the former considers the topic mixture coefficients as random variables of a Dirichlet distribution whereas the latter considers them as fixed unknown parameters. The selected low-level visual features and their combination used in [44, 45, 53] are the same as those in [52].

While topic models such as pLSA have been successfully adapted to solve the image annotation problem, the work presented in [35] takes one step backward to comparatively study the performances of latent semantic analysis (LSA) and pLSA. To represent each image, the bins of a local histogram and the keywords are collectively treated as the elements of a single vocabulary. For an image to be annotated, the dimensions of its feature vector corresponding to the keywords are zero. Projecting such a representation into the latent semantic space also causes problems, such as inaccurate image representation in the latent semantic space. Regarding the annotation methods for LSA and pLSA, the former is based on image-wise direct match in latent semantic space followed by keyword propagation and the latter is based on the *a posteriori* probability of a keyword given an image. Although the conclusion indicates that pLSA is inferior to LSA for image annotation, the performance evaluation scheme actually penalizes the ability of pLSA to learn the co-occurrence of words, as stated in [35]. In addition, pLSA can be used to construct hierarchical semantic models of an image collection, whereas LSA does not bear this ability. In terms of the combination of text and visual features, the image representation in [35] concatenating keywords and the bins of a local color histogram assigns equal importance to both modality, which is problematic as shown in [36]. To handle the inherently unequal contribution of the semantic meaning from the textual

and visual domains, two pLSA models are learned respectively using keywords and local visual features in [36], with the visual pLSA using the topic mixture coefficients resulting from learning the textual pLSA.

The above methods are primarily used for labeling images with keywords, which is the second task of automatic image annotation. In terms of region-based object annotation, the task was considered as a machine translation problem in [42]. The proposed approach is based on the language model developed in [54] and aims at learning the correspondence between object regions and keywords associated with the images, which are only weakly labeled. It is compared to the task of learning a lexicon from aligned bi-text for machine translation between documents of different languages. Since translation is inherently between two domains of discrete data, converting image features to discrete representation is necessary. To this end, the descriptors extracted from image segments resulting from Normalized-Cut [55] are vector quantized using the kMeans algorithm. The discrete representation is referred to as blob token. In essence, the statistical model is a table of conditional distributions $P(w|x)$, where w and x denote keywords and blob tokens. Given weakly labeled images, the correspondence between blob tokens and keywords is unobserved. By introducing a hidden variable for this relationship, the conditional probability table is estimated through an EM procedure, which can be directly used for annotating object regions of new images. A problem of this approach is that the segmentation is purely based on low-level visual features, which can not guarantee that all the resulting regions cover exactly meaningful objects, especially for cluttered scene.

Topic models have also been adopted for learning object categories from unlabeled image collections. In this case, the approaches have to learn high-level semantics of images completely from low-level visual features. Most of these methods use descriptors extracted from affine covariant regions [56, 57] or regions around scale-invariant salient

keypoints [58]. These descriptors are vector quantized with the kMeans algorithm in most of the existing works and the centers resulting from the clustering procedure are called visual words and used to form a visual codebook. As such, all the descriptors extracted from images are represented using elements in the codebook. This discrete data-based representation of images is normally called bag of visual words (BOVW), in correspondence to the bag of words representation of documents in the research area of document analysis. This is also the image representation framework adopted in the work presented in Chapter 5. The merits and drawbacks of the BOVW representation are:

- Feature vectors of a fixed length can be obtained regardless of the number of detected regions or keypoints;
- Efficient matching of images can be achieved;
- Discriminative power is unstable due to the dependence on the set of descriptors and the clustering method used to construct the codebook;
- The optimal number of visual words, i.e. the size of the codebook, varies from one dataset to another and is application-dependent as well.

In [59], pLSA and LDA were applied to discover object categories from unannotated image datasets. From each image, scale-invariant feature transformation (SIFT) descriptors [58] are calculated within the detected affine covariant regions, which are in turn used for calculating the BOVW image representation. To learn the topic distributions, an image is considered as a document which is composed of a set of visual words. The goal is to let the pLSA and LDA discover the set of topics, of which the distributions of visual words are consistent with the visual properties of various object categories existing in the image dataset. With the set of properly learned topics, both of the training images and

testing image can be classified into the object category with the maximum topic *a posteriori* probability, i.e. $P(O|d)$, where d denotes an image. Therefore, the second task of image annotation can be solved using such topic models without using text. Experiments for evaluating the effectiveness of the methods for this task were conducted progressively by increasing the number of categories, corresponding to the number of topics. It has been found that the topic models are able to automatically adapt the topic distributions of visual words to multiple meaningful object categories given an image dataset. However, problem arises when images containing different kinds of objects share similar visually coherent background, which occupies a large portion of each of these images. In this case, the dominant discriminative information will come from the background, which actually compromises the discriminative power of the model and the descriptors. By increasing the number of images of the similar background, the background will be identified as a separate topic. In addition, misclassification was also found to be related to cluttered background of images. In this case, an image includes many different sorts of objects and maybe multiple objects of the same kind. Experimental results showed that considerably increasing the number of topics will enhance the ability of the topic model to accommodate more object categories, resulting in restoring the discriminative power of the model. Like many other unsupervised learning techniques, the determination of the number of topics of the topic models is just another example of model order selection.

Noticing that the method in [59] regards each image as a document for learning topics, it is not difficult to understand that it results in the projection of the visual word distribution of an image into the space spanned by the topic distributions. Different from this perspective of modeling the image data, the method presented in [50] first segments each image using more than one segmentation algorithms and then treat the segments of all images as documents for topic modeling. As a result, this method projects the visual

word distribution of an image segment into the space spanned by a set of learned topics. By assuming that using multiple segmentation methods raises the chance of acquiring accurate extraction of semantically meaningful objects, the Kullback-Leibler divergence (KLD) between the visual word distribution of a topic and that of each segment is calculated, for each of the learned topics. Then, the method is claimed to be able to find not only the object categories using the BOVWs of segments but also the best segments of them.

Different from the topic models, which decompose the collection distribution or class distribution into several topic distributions, and the translation model, which directly estimate the probability of a word conditional on a region, several variants of relevance models were proposed in [37, 60, 61]. Relevance model was originally proposed as a language modeling approach to information retrieval [62, 63]. Essentially, a relevance model is the joint distribution of features from two different domains. When applied to image annotation, the distributions of textual keywords and visual features from image regions are jointly modeled, i.e. through $P(x, w)$. It can be used to annotate images and retrieve images from an unannotated database. In [37], a cross-media relevance model (CMRM) was presented. Image regions are represented as blobs in a similar way as [42]. Given the blobs $\{b_1, b_2, \ldots, b_m\}$ of an unannotated image, the joint distribution $P(w; b_1, b_2, \ldots, b_m)$ is calcuated through it decomposition into the conditional distributions of $\{w; b_1, b_2, \ldots, b_m\}$ given each of the images in the annotated training set, i.e. $P(w; b_1, b_2, \ldots, b_m | I_t)$, where I_t is the t-th training image. These conditional distributions are estimated by calculating the relative frequencies of the blobs and words within the training set. To deal with the sensitivity of CMRM to clustering errors, a continuous relevance model (CRM) was proposed in [60], where the visual features of regions are modeled using their PDF rather than PMF over the discrete blob representation. Es-

sentially, the CRM still decomposes the joint distribution of words and visual features into their distributions conditional on training images. In [61], it was pointed out that a multinomial distribution was not suitable for modeling the distribution of words in that the probability of a word in an image with shorter annotation is higher than that with longer annotation, which is supposed to be the same. A multiple-Bernoulli relevance model (MBRM) was introduced to handle this modeling issue with respect to the words.

Supervised Approaches

Most of the supervised approaches to image annotation fall into two categories, which are binary classification and multi-class classification. Let $\boldsymbol{C}$ denotes the set of classes. For each class $i \in \boldsymbol{C}$, the methods in the first group consider the samples of all the other classes j, where $j \in \boldsymbol{C}$ and $j \neq i$, as being in a single class, denoted $\bar{i}$. A decision boundary is learned between the classes i and $\bar{i}$. Hence, this type of approaches are also referred to as one-versus-all approaches. Eventually, a new image will go through a sequence of classifications using the classifiers for all classes in $\boldsymbol{C}$ and can be assigned multiple labels. If the problem is defined such that each sample, being an image or a region, belongs to one and only one class, the label of the binary classifier generating the greatest values is assigned to the sample. The second category of methods construct a multi-class model directly to annotate images or image regions. An image or region can also be assigned to multiple classes by choosing the top-k classes, of which the output values of the classifiers are greater than those of the others.

A typical example of the early works for supervised image annotation can be found in [31], where binary classification was employed to deal with scene classification for vacation images, such as indoor/outdoor, city/landscape, sunset/forrest, etc. These semantic categories are organized in a hierarchical structure identified by human subjects with

a small image collection. To be specific, an image is first classified into the category of outdoor or indoor. Then, if it is outdoor, the image is further classified as a city or landscape scene. Low-level visual color and texture features are extracted from sub-blocks of image tessellation, which are concatenated into a single vector representation. For each pair of scene types to be distinguished, the selection of features depends on the feature discriminative power with respect to the specific visual properties of the scene. The hierarchy of the semantic categories is different from that in [52]. In terms of classification, the distributions of visual features and text of all semantic classes are involved in [52] and hence is essentially a multi-class problem. It is the topic mixture coefficient, i.e. the *a posteriori* probability of topics given an image, determines the specific level of a concept. Being a multi-modal approach, [52] discover the topics which are salient joint distributions of visual features and text, which is suitable for a unsupervised learning.

In [47], binary classification was employed for multi-label image annotation. Associated with each assigned label of an image, there is also a factor indicating the confidence of annotating the image with the keyword. The vector consisting of the factors is used to match keyword-based queries for image retrieval. Color and texture descriptors are concatenated into a single vector to form the low-level visual feature representation of an image.

In contrast to [31, 47], a multi-class supervised framework based on a hierarchical mixture model was proposed in [38]. The method deals with weakly labeled data by taking the perspective of multiple instance learning [64, 65]. In their framework, an image is represented by a set of low-level visual feature vectors extracted from its regions. Each feature vector consists of discrete cosine transform coefficients of different color channels. A hierarchical mixture model is employed to estimate the probability density function (PDF) of each image, which is in turn used to estimate the PDF of a class. The

argument underlying the modeling technique is that the feature vectors belonging to a class of interest will become dominant when pooling together the feature vectors of all the training images of that class.

In [39,48], with the same problem formulation in terms of the image annotation task, two different models were proposed to characterize the class-dependent distributions of visual features. In [48], images are partitioned uniformly into patches, from which the statistics of the high-pass bands of wavelet transform are calculated. This feature extraction is performed at many different resolutions of an image and a 2-D hidden Markov model (HMM) was proposed to build the model of each class. As a supervised learning example incorporating multiple low-level visual descriptors, Li [39] proposed a framework in which an image is represented by two probability mass functions (PMFs), one for color and the other for texture. The average color and wavelet coefficients constitute the supports of the two PMFs respectively. Denoting the descriptors of two images by $x_i = (P_{i1}, P_{i2})$ and $x_j = (P_{j1}, P_{j2})$, where $P_{i1}, P_{i2}, P_{j1}, P_{j2}$ are PMFs, the distance between the two images, denoted D_{ij}, is defined as $\sum_{k=1}^{2} d^2(P_{ik}, P_{jk})$, where $d(P_{ik}, P_{jk})$ is the Mallows distance. The model for each class is built within the space of x_i, i.e. a set of PMFs.

Topic models have also been employed to annotate images using supervised learning framework. In such cases, each semantic class has its own mixture model constructed with the topic distributions of visual words. A representative piece of work is the hierarchical framework for scene classification proposed in [33]. Essentially, the scene classification is tackled by utilizing the LDA to model visual words, where the images belonging to a certain type of scene is regarded as a visual corpus. To emphasize the supervised nature of the approach, the distribution of the Dirichlet variable is formulated as a conditional distribution given the scene class, which is observed during the conceptual data generation

process. Different scene classes share not only the same visual word vocabulary but also the topic distributions of these visual words. In other words, various types of scene are considered to be mixtures of the same set of topics, the so-called themes in [33], with their own mixing weights. In this work, gray scale images were used for the experiments. Keypoint-based descriptors and those extracted from image patches were employed for image representation separately.

In contrast to the application of LDA to scene classification, pLSA has also been employed for handling the same task of image annotation. The major different between pLSA-based and LDA-based methods lies in the fact that pLSA can be used as an indexing method to produce the signature of each image within the latent semantic space because these signatures, in essence, constitute a subset of the parameters of the model. On the other hand, LDA is a relatively compact model compared with pLSA in that it considers the document signatures, i.e. topic mixture coefficients, as random variables, following a Dirichlet distribution, instead of fixed unknown parameters as in the case of pLSA. Therefore, LDA does not produce image signatures in the topic space but merely the parameters of the distribution of them. Taking advantage of this characteristic of pLSA, the image signatures in the latent semantic space, i.e. topic space, can be used to build discriminative classifiers therein. This is just the line of thought of the work presented in [32], where the image signatures were used to learn a set of SVM's for recognizing the scene category given a new image. SIFT descriptors were extracted from different color channels followed by vector concatenation to build a single vector representation of each keypoint location.

2.3.2 Image Annotation Using Context-Aware Models

At the same time, studies in the field of psychology [66,67] and neuroscience [68] have been conducted to investigate the advantage of incorporating contextual information into the process of scene perception. If available, contextual information can be utilized to refine the distribution of the object categories because objects do not appear simultaneously in a random fashion. Each object category has it own frequently co-existing object categories and they are often present according to a fixed spatial relation. In general, there are five rules that govern how objects exist in a real world scene, which are: 1) interposition (background is partially obscured by foreground), 2) support (objects usually stay on surface), 3) probability (an object may often appear in one scene but not others, 4) position (an object of a kind has its frequent location or area in a scene), and 5) familiar size (the size of an object relative to those of other objects in the same scene). In this thesis, the contextual information refers to the statistical dependence and spatial relation among different image sub-structures, i.e. regions or patches, and various object categories. Usually, the dependence across multiple image sub-structures is introduced by taking into account the dependence among the object categories. Moreover, scene category, if used as a contextual information, is regarded as the condition to distinguish the distribution of visual features of one type of scene from that of another. Interdependence among scene categories is rare. The concept of scene in the formulation of the probability rule can also be defined by the existence of object categories other than the object category of interest.

A large amount of works on machine-based visual recognition have been proposed to address the problem of incorporating contextual information into the classification process. These approaches are different in terms of the specific nature of the contextu-

al information. The first kind of contextual information is defined as the global visual property characterizing the category of the scene. Studies in the research field of psychology [69–71] has shown that visual perception works in a coarse-to-fine order in processing complex scene. To recognize the scene category, detection and recognition of individual objects in a scene are not necessary. In fact, the study also showed that the recognition of the scene category can actually provide auxiliary information for further recognizing the categories of individual objects. Following this implication, research in the area of visual recognition has exploited global visual statistics as the contextual information to boost the performance of object recognition [41, 72]. The second type of contextual information is defined as the interaction between object categories, which can be further divided into two sub-categories, which are statistical dependence, as in [43, 73–75], and spatial relation among different objects, such as [76], respectively.

In [41], the context in the form global scene visual properties was introduced into the Bayesian object category recognition, i.e. the contextual features appear as one of the conditional variables in $P(O|x, c)$, where x and c denote object visual features and contextual features and O denotes the object category. This *a posteriori* probability can be decomposed into two factors, with the second factor being $P(O|c)$. Based on the research in [77], the contextual information refining the object category distribution can be regarded as a single object category and represented in a low-dimensional space. According to [78, 79], suitable candidates for playing the role of context include: 1) Statistics of structural elements, i.e. textures, 2) spatial organization, and 3) color distribution. In this work, spatially localized structural information was extracted by applying a set of oriented bandpass filters. The output of these filters were further projected into a lower dimensional space using the principal component analysis (PCA). Based on this global contextual feature, the graphical model proposed in [80] combines the estimated

a posteriori object category probability $P(O|c)$ and a bottom-up local object detectors. The resulting model can solve object detection and scene classification at the same time.

To incorporate statistical dependence among the object categories of neighboring blobs into the annotation process using the machine translation model [42], Markov random field (MRF) was employed in [43]. MRF is a generative framework modeling the joint distribution of an image and its labels, which has been extensively used for low-level image processing problems [81,82]. In the experiments of [43], image segmentation using Normalized Cut and uniform image partitioning into non-overlapping grids were considered. For each image, the hidden state of a blob is constrained to take a value from the set of keywords in its annotation. Because both blobs and words belong to a finite set in their respective domains, two tables of potential functions can be defined, characterizing the possibility of assigning a word to a blob and that of labeling two adjacent blobs with a particular pair of words, respectively. Given an image, the complete likelihood can be evaluated using these two tables for learning the model using an EM procedure. The loopy belief propagation [83] was used to evaluate the *a posteriori* probability of the hidden variables in the E-step. For the M-step, the complete likelihood was decoupled into pseudo-likelihood as in [84], which was in turn optimized using the iterative scaling [85]. Results showed that the approach is suitable for smoothing the label assignment and has the potential to grouping over-segmented regions for the purpose of segmentation.

To make the model learning computationally tractable, generative context-aware approaches, such as MRF, make restrictive assumption that the observations are statistically independent given the labels. To relax this assumption and make the model capture any arbitrary dependence across the observations, discriminative frameworks were proposed. Instead of modeling $P(\boldsymbol{X}, \boldsymbol{Y})$ as in generative approaches, where $\boldsymbol{X}$ and $\boldsymbol{Y}$ denote labels and observations, discriminative models directly characterize $P(\boldsymbol{X}|\boldsymbol{Y})$ for the purpose

of classification. Conditional random field (CRF), originally proposed for labeling 1-D sequential data [86], is a representative model for context-aware modeling in a discriminative fashion. Due to its superior performance to generative approaches, many variants of CRFs have been proposed, which were extensively used for modeling 2-D image data in existing works. These approaches mainly differ from each other in terms of: 1) the scope of the image over which label interaction is defined [87], 2) the specific definition on the association potential and interaction potential [74, 88], and 3) The modeling target of the labels in terms of the level of granularity of the image structure [89].

In [88], a discriminative random field (DRF) model was proposed. Not only the top level framework, i.e. the CRF, is discriminative in nature, its local potentials, i.e. the association potential and the interaction potential, are also discriminative models. Generalized learning models (GLM) are used for defining such potentials followed by converting the output of these GLMs to probabilities using a logistic function. Using iterative conditional mode for inference, DRF outperforms MRF in the application to the detection of man-made structures in images. This binary classification framework for object detection was generalized in [75] to solve multi-class problems, and hence can recognize object categories within an image.

In [87], a CRF modeling object category interaction at different scales was proposed. At each location within an image, where a category label is to be assigned, there are multiple interaction potentials defined over the labels of neighboring locations. These potentials are distinguished from one another by the scope of the neighborhood. Three different selected ranges are pixel, region, and global range. According to the parametric form of CRF, the information captured at different scales are combined multiplicatively.

Instead of modeling the statistical dependence across multiple image segments or blobs as in [43], the dependence among the visual features extracted from image regions

around salient keypoints were taken into consideration using CRF in [89]. These regions are detected using the keypoint detection algorithm in [58] and considered as parts of objects. The model was named hidden conditional random field (HCRF) because a hidden layer of variables are introduced to represent the part labels of an object. Part labels represent finer details compared with the object categories. In addition to the observed visual features from regions around keypoints, the object category contained by an image is also observed. Therefore, it is essentially a supervised classification method with inter-dependent part labels. In fact, the concept of object part is essentially the same as that of a topic in topic modeling. The topology of the CRF was defined to be a tree structure such that exact inference algorithm, such as belief propagation [90], can be exploited to estimate the parameters.

In [74], without specifying a neighborhood over nearby locations of an image to be labeled, a fully connected graph was employed for the interaction potential defined over labels. The estimation of the parameters of this CRF is not performed through the maximum likelihood of exponential models but via the estimation of the parameters of another distribution, of which the parameters are the potential functions of the CRF for the object category recognition task. The model used to learn the parameters is the joint distribution of a set of binary random variables, each of which corresponds to one object category and takes on the value of 1 if the image contains the corresponding object and 0 otherwise.

2.4 Summary

Through the literature review presented in this chapter, it is not difficult to realize the following implications to the methodology for improving the performance of image

retrieval and annotation systems.

- For image retrieval, the combination of STRF and LTRF is a promising approach to utilize the knowledge acquired from both on-going HMI and past HMI, which compose the set of primary information sources to train an effective retrieval system. To this end, a mathematically justifiable framework is desirable. The system with both STRF and LTRF enabled should be functional even no knowledge is available for training the LTRF. In addition, it should be able to gradually accumulate necessary information in order to incrementally upgrade the system for LTRF in a way similar to how human memory works. More interestingly, considering the fact that there are many objects which can generate characteristic sound, this can also be utilized to compensate for the inefficiency of visual information in terms of handling the semantic gap.

- For image annotation, much effort has been taken to construct sophisticated statistical models to improve the classification performance. While being aware of the importance of utilizing various visual features to represent images, most of the existing approaches either adopt the vector concatenation of low-level visual descriptors or the concatenation of the BOVW representation, if the quantization of visual features is used. Moreover, some statistical models, such as the topic models, characterize the distributions of multiple features via factorizing their joint distribution into topic distributions in respective feature domains; however, this sort of approaches have only been employed to handle the modeling of text and visual features. Since different types of low-level visual features have much stronger correspondence in terms of their spatial locations, a novel modeling technique is needed to characterize the joint distributions of the visual features.

In the subsequent chapters, two frameworks, namely a Bayesian framework and a multi-feature probabilistic semantic analysis (MF-pLSA) framework, are presented. The former is applied to both image retrieval and annotation problems and the latter is used to solve image annotation by using multiple visual features more effectively.

Chapter 3

A Bayesian Framework for Image Annotation and Retrieval

3.1 Introduction

Based on the background introduction and literature review in previous chapters, the framework and its related experiments to be presented in this chapter are primarily focused on the integration of low-level visual content and contextual information. As mentioned earlier, the contextual information in the presented work refers to the statistical correlation across multiple entities, where the entities are images for the application to image retrieval and different semantic classes for the application to image annotation. For image annotation, we are seeking an efficient approach to integrating visual content and context. For image retrieval, an approach which can utilize both STRF and LTRF in a unified framework is desired, which are based on content and context as well. Motivated by such goals, a Bayesian framework is developed in which the *a priori* probability, learned through a maximum entropy algorithm, represents the contextual information

and the likelihood evaluation corresponds to the visual content analysis. In addition, the framework can utilize various models for the purpose of content analysis. In both application scenarios, these two components refine each others' evaluation of the similarity between images or an image and a semantic concept. Principally, the underlying rationale of the integration is that the online observation of visual content refines the *a priori* information encoded in the context model, especially when there is not sufficient high-level knowledge, whereas the contextual information can be used to bridge, to some extent, the semantic gap associated with the low-level visual features.

3.2 The Framework for Integrating Visual Content and Context

3.2.1 The Integration of Content and Context

The notation used throughout the elaboration of the Bayesian framework is introduced first. Assuming the feature of an observation is a vector in a d-dimensional feature space, it is denoted x, where $x \in R^d$. Let $\boldsymbol{W}$ represent the set of class labels and $\boldsymbol{W} = \{1, 2, \ldots, W\}$, where W is the number of classes. The class label of a particular observation is denoted ω, where $\omega \in \boldsymbol{W}$. Based on the maximum *a posteriori* probability (MAP) criterion which minimizes the classification error, the true class label is estimated with

$$\hat{\omega} = \arg\max_{\omega \in \boldsymbol{W}} P(\omega | x, \boldsymbol{I}), \tag{3.1}$$

where $\hat{\omega}$ is the estimate of ω. In applications such as image retrieval, where no decision with the notion of the most probable estimate is made, the *a posteriori* probability can

be used as a relevance score to rank the classes. In the literature, $\boldsymbol{I}$ is normally referred to as the background information, which exists with a well-formulated problem. In the context of the subsequent description, it represents a set of indexes of either semantic classes or query images, depending on the nature of the application. Therefore, $\boldsymbol{I}$ can be defined as $\{I_i | i = 1, 2, \ldots, |I|\}$, where $|I|$ is the number of indexes of $\boldsymbol{I}$. How to acquire this piece of information and how it contributes to the accomplishment of a given task will be elaborated in next section. Using the Bayes' theorem, the *a posteriori* probability can be written as

$$P(\omega|x, \boldsymbol{I}) \propto p(x|\omega, \boldsymbol{I})P(\omega|\boldsymbol{I}), \tag{3.2}$$

with the equality replaced by the proportionality due to the unimportance of the probability density function (PDF) of an observation, i.e. $P(x|\boldsymbol{I})$, when the theorem is employed to solve a classification problem. Based on the meaning of the background information $\boldsymbol{I}$, we can assume the conditional independence between the observation x and $\boldsymbol{I}$ given the class label of the observation, i.e. $x \perp \boldsymbol{I}|\omega$. Therefore, the *a posteriori* probability in (3.2) can be calculated through

$$P(\omega|x, \boldsymbol{I}) \propto p(x|\omega)P(\omega|\boldsymbol{I}). \tag{3.3}$$

The first term on the right-hand side of (3.3) is the PDF of the feature vector of the class ω, which is considered as the content model characterizing the visual properties of that class. Given the afore-mentioned definition of $\boldsymbol{I}$, the second term is essentially a distribution of one class or candidate image, say ω, conditional on a set of other classes or query images, collectively represented by $\boldsymbol{I}$. This is exactly the contextual information that characterizes the statistical relation between different classes or images. It will be

shown that such contextual information can be learned from weakly labeled images for automatic annotation and past user feedback for image retrieval. According to (3.3), the content and contextual information are integrated through the decision-level fusion in a multiplicative fashion. Before proceeding to the discussion on the retrieval and annotation frameworks, the learning algorithms associated with the content and context components of the Bayesian framework are introduced here.

3.2.2 Learning the Content Model

The visual content model of a certain semantic class, e.g. ω, is the parametric form of the distribution of the visual features of that class. The parameters of the model are adapted to a given set of training data of class ω through a supervised learning procedure. Since a visual content model plays the role of evaluating the likelihood of a visual feature with respect to a certain class, any parametric or non-parametric model can be applied, as long as it can quantitatively measure the degree of consistency of a visual feature with it. We select the support vector machine (SVM) as the key component of the content model to evaluate the likelihood, considering its high discriminative power for many applications. In the application to image retrieval, L1-norm is also employed in addition to SVM for calculating the likelihood using the content model, which shows the flexibility of the proposed Bayesian framework as mentioned before. At the same time, it should be noted that the formulation of the Bayesian framework requires that the output of the visual content model comply with the definition of a PDF. To this end, we employ the exponential function, i.e. $h(s) = \exp(s), s \in R$, to convert the discriminant function of SVM into a PDF. The selection of the above exponential function is based on the following consideration. First, it is monotonically increasing, resulting in the

preservation of the physical interpretation of the algebraic distance between a sample and the decision boundary. Second, it is positive. Since the total integral of a function must be equal to unity, appropriate normalization is necessary. Finally, representing the discriminant function of SVM corresponding to the ω-th class as $f_\omega(x)$ and substituting it for the variable s in the exponential function followed by normalization, we obtain

$$p(x|\omega) = \frac{1}{A}\exp(f_\omega(x)), \tag{3.4}$$

where $A = \int \exp(f_\omega(x))dx$. When the likelihood is calculated using the L1-norm, the corresponding negative distance function should be substituted into the exponential function because the similarity is a decreasing function of the distance between features. More details regarding this issue will be revealed in the Section 3.5.

3.2.3 Learning the Context Model

In this part, our objective is to calculate the $P(\omega|\boldsymbol{I})$ in (3.3), which is the contextual information about ω inferred based on the $\boldsymbol{I}$. Without $\boldsymbol{I}$, the probability mass of ω is uniformly distributed over the class ensemble $\boldsymbol{W}$ without $\boldsymbol{I}$. Due to the statistical dependence across different classes, however, the distribution of ω conditional on $\boldsymbol{I}$ will deviate from the uniform distribution once $\boldsymbol{I}$ is available. As a result, the classes that are more strongly correlated with $\boldsymbol{I}$ have higher probabilities than the others do. Since the problem is essentially the estimation of a conditional probability mass function (PMF), a typical train of thought leads to the conventional approach that calculates the conditional probability through $P(\omega|\boldsymbol{I}) = P(\omega, \boldsymbol{I})/P(\boldsymbol{I})$, for which we need a set of training samples belonging to the cartesian product of $|\boldsymbol{I}|+1$ $\boldsymbol{W}$'s. Regardless of the approach to estimating $P(\omega, \boldsymbol{I})$ and $P(\boldsymbol{I})$, there are two problems with above estimation on $P(\omega|\boldsymbol{I})$.

First, the background information $\boldsymbol{I}$ may include different numbers of indexes, which requires separate estimation of the model for different sizes of $\boldsymbol{I}$. Second, when collecting training data, we can not guarantee enough or even available samples for a certain configuration of ω and $\boldsymbol{I}$, where by configuration it means a particular instance of the number of random variables of $\omega \cup \boldsymbol{I}$ and their values. We propose the following way of modeling the contextual information.

To deal with the estimation on the context model efficiently, we propose to approximate the $P(\omega|\boldsymbol{I})$ using a distribution of a set of binary random variables estimated based on the maximum entropy (MaxEnt) principle. In this approach, an image is represented using a W-dimensional vector of binary random variables, denoted $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_W)^T$, where the value of each variable Y_ω is defined by

$$Y_\omega = \begin{cases} 1, & \text{if an image is labeled with } \omega \text{ or if image } \omega \text{ is relevant to a query,} \\ 0, & \text{otheriwse.} \end{cases} \tag{3.5}$$

Instead of being from the cartesian product of $|\boldsymbol{I}| + 1$ $\boldsymbol{W}$'s, the data utilized by the proposed context modeling procedure belong to the set of vertices of a W-dimensional hypercube. Given a set of T training samples, denoted $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_T$, we can estimate the $P(\boldsymbol{Y})$ and then calculate the conditional probability $P(Y_\omega|Y_{I_1}, Y_{I_2}, \ldots, Y_{I_{|I|}})$, which is represented as $P(Y_\omega|Y_{\boldsymbol{I}})$ in what follows. To approximate the $P(\omega|\boldsymbol{I})$ in (3.3), we use

$$P(\omega|\boldsymbol{I}) = \frac{P(Y_\omega|Y_{\boldsymbol{I}})}{\sum_{v=1}^{W} P(Y_v|Y_{\boldsymbol{I}})}. \tag{3.6}$$

As the size of the concept ensemble, i.e. W, grows, the computational intensity of the calculation of $P(Y_\omega|Y_{\boldsymbol{I}})$ increases exponentially. Therefore, it would be more efficient if we can directly estimate $P(Y_\omega|Y_{\boldsymbol{I}})$ based on a set of training samples. To this end, we employ

the MaxEnt approach proposed in [91], which estimates a conditional distribution by maximizing its Rényi entropy. Essentially, the MaxEnt principle states that the optimal model should only respect a certain set of statistics induced from a given training set and otherwise be as uniform as possible. The MaxEnt approach employed in our study searches for the conditional distribution $P(Y_\omega|Y_{\boldsymbol{I}})$, with the maximum entropy, among all the distributions which are consistent with a set of statistics extracted from the training samples. Therefore, it can be considered as constrained optimization, which is formulated as

$$\max_{P(Y_\omega|Y_{\boldsymbol{I}})\in[0,1]} -\sum_{y_\omega, y_{\boldsymbol{I}}} \hat{P}(Y_{\boldsymbol{I}} = y_{\boldsymbol{I}})P(Y_\omega = y_\omega|Y_{\boldsymbol{I}} = y_{\boldsymbol{I}})^2,$$

subject to:

$$\frac{\sum_{y_{\boldsymbol{I}}} \hat{P}(Y_{\boldsymbol{I}} = y_{\boldsymbol{I}})P(Y_\omega = y_\omega|Y_{\boldsymbol{I}} = y_{\boldsymbol{I}})f_k}{\hat{P}(f_k)} = \hat{P}(f_\omega|f_k), k \in \{0\} \cup \boldsymbol{I},$$

where $\omega \in \boldsymbol{W}$ and $\omega \notin \boldsymbol{I}$ because $P(Y_\omega = 1|Y_I = 1) \equiv 1$ for $\omega \in \boldsymbol{I}$. In addition, $\hat{P}(\cdot)$ represents the empirical probabilities directly estimated from the training samples, $f_\omega = Y_\omega$ and $f_k = Y_k$ when $k \neq 0$ and $f_k = 1$ otherwise. Using a matrix-based representation, solving the above optimization leads to the result that

$$\boldsymbol{P} = \boldsymbol{M} \times \boldsymbol{N}^{-1} \times \boldsymbol{f}, \tag{3.7}$$

where

$$\boldsymbol{P} = \Big(P(Y_{a_1}|Y_{\boldsymbol{I}}), P(Y_{a_2}|Y_{\boldsymbol{I}}), \ldots, P(Y_{a_{|\boldsymbol{W}/\boldsymbol{I}|}}|Y_{\boldsymbol{I}})\Big)^T, \tag{3.8}$$

$$\boldsymbol{M} = \begin{pmatrix} \hat{P}(f_{a_1}|f_0) & \hat{P}(f_{a_1}|f_{I_1}) & ... & \hat{P}(f_{a_1}|f_{I_{|\boldsymbol{I}|}}) \\ \hat{P}(f_{a_2}|f_0) & \hat{P}(f_{a_2}|f_{I_1}) & ... & \hat{P}(f_{a_2}|f_{I_{|\boldsymbol{I}|}}) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{P}(f_{a_{|\boldsymbol{W}/\boldsymbol{I}|}}|f_0) & \hat{P}(f_{a_{|\boldsymbol{W}/\boldsymbol{I}|}}|f_{I_1}) & ... & \hat{P}(f_{a_{|\boldsymbol{W}/\boldsymbol{I}|}}|f_{I_{|\boldsymbol{I}|}}) \end{pmatrix},$$

$$\boldsymbol{N} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ \hat{P}(f_{I_1}|f_0) & 1 & \dots & \hat{P}(f_{I_1}|f_{I_{|\boldsymbol{I}|}}) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{P}(f_{I_{|\boldsymbol{I}|}}|f_0) & \hat{P}(f_{I_{|\boldsymbol{I}|}}|f_{I_1}) & \dots & 1 \end{pmatrix},$$

and

$$\boldsymbol{f} = \left(f_0, f_{I_1}, \dots, f_{I_{|\boldsymbol{I}|}}\right)^T, \tag{3.9}$$

where $\boldsymbol{W}/\boldsymbol{I} = \{a_1, a_2, ..., a_{|\boldsymbol{W}/\boldsymbol{I}|}\}$.

3.3 The Application to Image Annotation

3.3.1 Overview

In terms of improving the performance of annotation, our work is motivated by the following analysis. Among the information at our disposal, the content and context are two critical resources. In the presented work, the former refers to the low-level visual features extracted from image regions obtained through image segmentation and the latter can be interpreted as the co-occurrence across different real world object categories in a probabilistic sense or in other words, the statistical dependency across different object categories. In the rest of this chapter, the term semantic concept, semantic class and

object category are used interchangeably. In particular, the contextual information plays a fairly critical role in the sense of incorporating high-level knowledge into the underlying classification process. To be specific, one aspect of the image annotation task in question that distinguishes it from conventional pattern classification tasks is that samples are presented to the system in a batch, typically in the form of a set of regions/segments constituting an image, in which the states of nature of the samples depend on each other statistically. For example, suppose we intend to annotate an image containing two regions, which respectively correspond to an animal and water. Based on our common knowledge, an intuitive inference with respect to the animal is that its probability of being an aquatic is higher than that of being a terricolous one. As mentioned earlier, this inter-dependence across the concepts associated with different regions of an image can be referred to as contextual information, which can be used as auxiliary information to acquire performance improvement. This is also a particular example of the results from the psychological study in [66, 67] reviewed in the previous chapter.

The annotation framework was developed through two stages. In the first stage, a semi-automatic mechanism was designed, in which a limited amount of user-machine interaction (UMI) is needed to produce the contextual information for the Bayesian integration of content and context. In the second phase, the framework was upgraded such that image search was leveraged to discard the necessity of UMI during the annotation process. The details of the second stage is elaborated in a separate section. With a set of pre-annotated images, the visual features are used to train an ensemble of probabilistic classifiers, each of which corresponds to a semantic concept, and the textual annotation is used to generate a probabilistic model encoding the contextual knowledge. The outputs of the content and context models are combined based on the Bayes' theorem discussed in the previous section. The underlying rationale is that the context and

content models make contribution to the final decision before and after examining the visual appearance of objects, which naturally follows the way the Bayesian decision theory treats information. Since the context model is estimated through collaborative filtering based on the maximum entropy (MaxEnt) principle, the proposed method is referred to as collaborative Bayesian image annotation (CBIA).

Meanwhile, it is worthwhile to mention that the CBIA framework is developed to solve the third task of automatic image annotation, i.e. the region-based object annotation. We consider the construction of the association between image regions and a set of pre-defined concepts, represented using keywords; that is, given an image, a concept is assigned by a machine to each of the regions resulting from image segmentation. Compared with the image-oriented annotation, which assigns one or more keywords to an entire image, the advantage of this region-oriented annotation is two-fold. From the view point of the system, local features extracted from a certain region are better representations of the visual property of an object or part of an object, resulting in better performance of classification. From the perspective of the users, the establishment of the relation between each region of a new image and a concept generates more useful information than the image-oriented annotation. For example, the region-oriented annotation figures out not only the existence of an object but also the location of it, which can be further exploited for the purpose of either knowledge acquisition from a user's point of view or location-dependent classification from a machine's point of view.

3.3.2 Visual Content and Context Analysis

As illustrated in Fig. 3.1, the framework is composed of a set of modules with different functionalities, which can be divided into offline and online sub-system. The operation

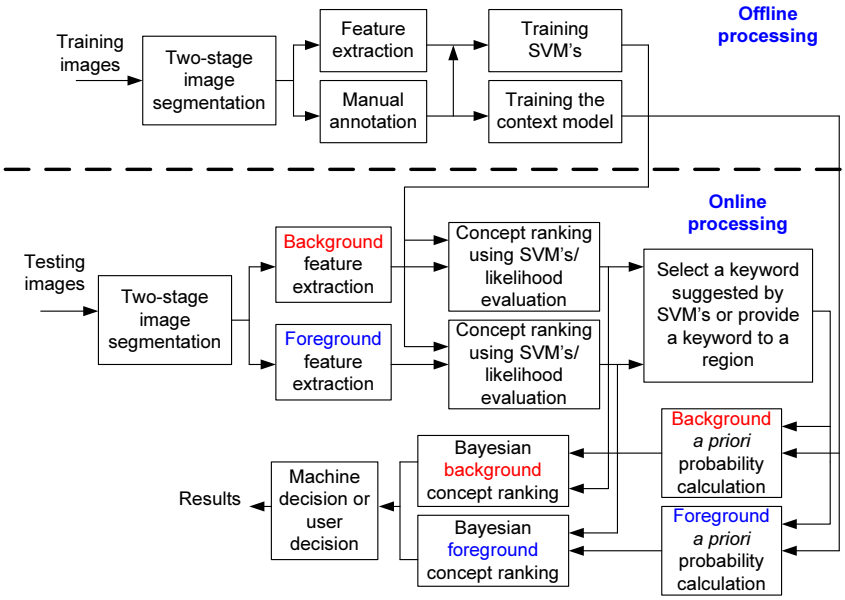
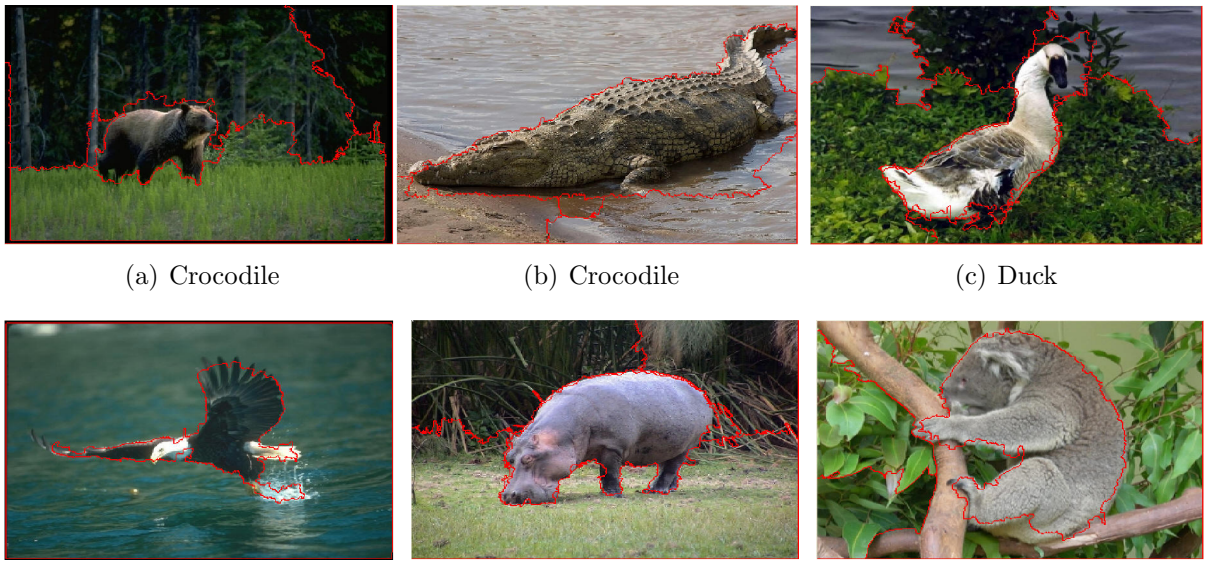


Figure 3.1: The system block diagram of the CBIA framework.

of the offline sub-system is primarily responsible for system training. A set of training images are processed using a two-stage image segmentation, as described in [92]. The objective of using this two-stage image segmentation is to classify each segment into foreground or background class in addition to partition an image into several visually coherent regions. Simply put, the benefit of the two-stage image segmentation lies in the reduction of the number of candidate semantic classes when a decision needs to be made, leading to the alleviation of the semantic gap to some extent. Accordingly, the concept ensemble is divided into two mutually exclusive and exhaustive subsets, one composed of the foreground concepts and the other made up of the background ones. Denote the two groups by $\mathbf{W}_F$ and $\mathbf{W}_B$, the decision rule in (3.3) can be re-written as

$$\hat{\omega} = \begin{cases} \arg\max_{\omega \in \mathbf{W}_F} P(\omega|x, \boldsymbol{I}) & \text{if } x \text{ is extracted from a foreground segment} \\ \arg\max_{\omega \in \mathbf{W}_B} P(\omega|x, \boldsymbol{I}) & \text{if } x \text{ is extracted from a background segment} \end{cases}, \quad (3.10)$$



(a) Crocodile (b) Crocodile (c) Duck

(d) Crocodile (e) Hippo (f) Kola bear

Figure 3.2: Examples of segmentation results.

where $P(\omega|x, \boldsymbol{I})$ is obtained using (3.3).

It has been recognized that automatic yet accurate algorithms for extracting meaningful objects from images are difficult to achieve in practice. Usually, unless the visual properties on the two sides of the boundary of an object are considerably different, the segmentation results in a few image regions, none of which covers exactly the entire region of a single object. Another common observation is that several objects belonging to different semantic categories are grouped into a single image region. On the other hand, interactive image segmentation may achieve improved segmentation performance by incorporating human knowledge into the segmentation process. In view of this, the interactive GrabCut [93] is used to perform foreground and background segmentation in the first step. To minimize the UMI, Grabcut only needs users to drag a rectangular box around the desired foreground. Second, an automatic image segmentation method is used to further partition the background into several homogenous regions (background

objects). To this end, the mean shift algorithm [94] is employed as the automatic image segmentation method. Nonetheless, precise image segmentation is not required, as we rely little on the shape of a certain object to recognize its semantic content in the presented framework. To visualize the results of in the image segmentation step, some selected examples are shown in Fig. 3.2. After image segmentation, color and texture descriptors are extracted from the regions. Color moments [95] are used as the color feature by extracting the mean and standard deviation from each channel of LUV color space. Wavelet moments [96] are used as the texture feature, by applying a three-level wavelet decomposition on the image followed by extracting the mean and standard deviation of the transform coefficients. To encode the content information, the set of low-level features extracted from the image regions are used to train a set of SVMs via supervised learning, which correspond to the semantic concepts in our pre-defined vocabulary. Meanwhile, the keywords associated with the annotated images are used to build a statistical model, characterizing the contextual information.

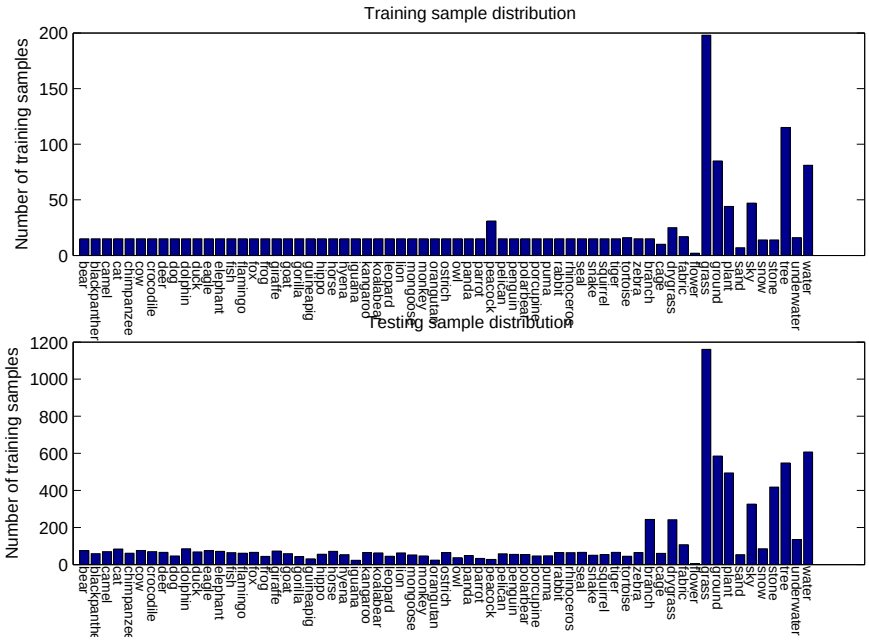
In the on-line mode, an image undergoes the same procedure of content analysis as those training images at first. Afterwards, the concepts are ranked for each of the image regions using the trained SVMs. At this point, a user may check the ranked lists of concepts and select a correct one appearing in one of the ranked lists, or provide the concept of a region if none of the regions has a correctly suggested one. It is this input from the user that is considered as the background information $\boldsymbol{I}$ in the formulation of the Bayesian framework. The user's feedback will serve as the input of the context model, which generates the probability, expressed in (3.6), of the appearance of other concepts given the one provided by the user. Finally, with the modification of the content model and the context model elaborated previously, the two types of information are integrated through the Bayesian framework.

Figure 3.3: The sample images in the database employed for performance evaluation.

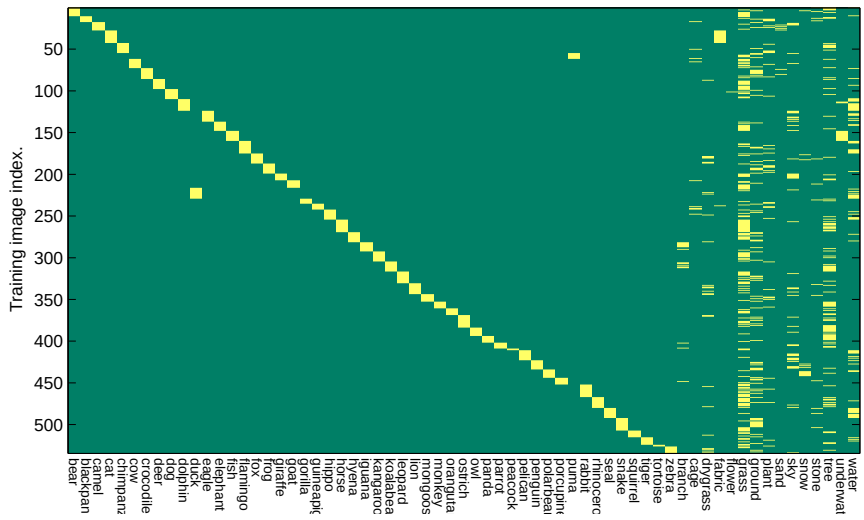
3.3.3 Experiments

Database and Concept Vocabulary

To evaluate the performance of the proposed framework and compare it with several other alternative approaches, we used an image database consisting of 5000 images featuring 50 different categories of animals, which were collected from Flickr™ and Google™. In each category, there are 100 images. It covers a wide variety of species of animals, with some examples shown in Fig. 3.3. Based on the scope of the semantic content of the database, we defined a concept vocabulary including 65 keywords. In addition to the concept corresponding to the 50 kinds of animals, there are 15 concepts representing the real-world objects that possibly appear in the environment where the animals live. Therefore, $W = 65$. The specific keywords can be found in Table. 3.1.



(a) The number of samples for training the SVMs and the number of testing samples in each semantic class.



(b) The training samples for the concept network.

Figure 3.4: The information on the training and testing sets in terms of the number of image segments.

Table 3.1: The foreground and background concepts considered in the presented study.

Foreground	**concept**			
1. bear	2. black panther	3. camel	4. cat	5. chimpanzee
6. cow	7. crocodile	8. deer	9. dog	10. dolphin
11. duck	12. eagle	13. elephant	14. fish	15. flamingo
16. fox	17. frog	18. giraffe	19. goat	20. gorilla
21. guinea pig	22. hippo	23. horse	24. hyena	25. iguana
26. kangaroo	27. koala bear	28. leopard	29. lion	30. mongoose
31. monkey	32. orangutan	33. ostrich	34. owl	35. panda
36. parrot	37. peacock	38. pelican	39. penguin	40. polar bear
41. porcupine	42. puma	43. rabbit	44. rhinoceros	45. seal
46. snake	47. squirrel	48. tiger	49. tortoise	50. zebra
Background	**concept**			
51. branch	52. cage	53. dry grass	54. fabric	55. flower
56. grass	57. ground	58. plant	59. sand	60. sky
61. snow	62. stone	63. tree	64. underwater	65. water

Table 3.2: The information on the training and testing sets in terms of the number of images.

Data set	Number of images
Training set for content model	750
Training set for context model	534
Testing set	2855

Training and Testing Sets

After the preprocessing, each region of an image is manually assigned a keyword, selected from the concept vocabulary. With no special consideration, we selected 750 images, with 15 from each of the 50 animal categories, to train a set of SVMs for the pre-defined concepts. For each semantic category, the training set is composed of the visual features of the image segments containing the objects of the semantic category. Since the essential

notion of context in the presented work is the possibility of the co-occurrence of the objects of different semantic categories, to model the contextual information and test the proposed framework, only images containing more than one semantic concepts can be used. According to this requirement, only 534 of 750 images used to train the SVMs are useful for training the context model. In addition, other than the images for training the SVMs and the context model, there are 2855 images available for performance evaluation and comparison in the database. The above usage of the images for training and testing purposes is summarized in Table. 3.2. Since the annotation is considered as a pattern classification problem in which each sample, for either training or testing, is the feature vector of an image segment, it is more informative to look into the sizes of training and testing sets in terms of the number of image segments. To this end, the number of image segments belonging to each semantic category is shown in Fig. 3.4(a). In addition, it has been calculated that there are in total 7927 segments used as testing samples with 2.78 segments per image on average. It is due to the fact that images containing different kinds of animals may include the same type of background objects that there are significantly more segments of background concepts than those of foreground concepts. It should also be noted that, unlike the case of training SVMs, the concept-to-segment alignment is not needed for learning the context model. The co-occurrence of different semantic concepts within the 534 images for training the concept network is illustrated in Fig. 3.4(b).

Performance Evaluation Criteria

First, as a general criterion for evaluating all statistical pattern classification systems, the average classification accuracy can be employed to measure the overall performance of the proposed annotation framework. In the context of the presented work, the average classification accuracy, denoted P_{avg}, is defined as, given an image segment, regardless

of what the true semantic category is, the probability of assigning a keyword that is consistent with the ground truth. To approximate P_{avg}, the ratio of N_{CS}, the number of correctly classified segments, to N_{TS}, the total number of testing segments is evaluated, i.e. $P_{avg} \simeq N_{CS}/N_{TS}$, where $N_{TS} = 7927$. The above measure is also extended in the following way, corresponding to the second case mentioned earlier. When there is a large concept vocabulary, if a machine can suggest a relatively small set of relevant keywords containing the correct one and leave the final decision to human users, it is still helpful in terms of annotation efficiency. To study the performance under such a circumstance, we employ $P_{avg}(k) \simeq N_{CS}(k)/N_{GS}$, where $N_{CS}(k)$ is the number of correctly classified segments by examining the top k concepts on the ranked list. In this case, a segment is correctly annotated as long as the actual semantic concept appears within the top k ones ranked using various methods. It can be seen that, when $k = 1$, it is just the performance of the machine-based decision.

Second, as the most effective visual features for characterizing different semantic content are different as well, with the same low-level representation applied to all images in our study, the severity of the semantic gap associate with each category is variable. Hence, the knowledge that which semantic class has the lowest accuracy is valuable. In addition, considering that the annotation task under consideration is a multi-class classification problem, in which there are $W - 1$ types of error for each class, further investigation on the probability of different errors also discover suggestive information on how to bridge the semantic gap. To study the performance of the proposed system from these perspectives, we employ the confusion matrix. Denoted by R, the element on the i-th row and j-th column is defined as $R_{i,j} = N_{i,j}/N_j$, i.e. the ratio of the number of segments belonging to category j and classified into category i, to the total number of segments belonging to category j, where $i, j \in \{1, 2, ..., W\}$. This ratio can be interpreted

as the approximation of the classification accuracy of each category, when $i = j$, and the probabilities of different types of classification errors within the j-th category, when $i \neq j$. It facilitates the investigation of the semantic gap specific to each pair of concepts, provided that the number of testing samples is sufficiently large. It should be noted that, as long as being measured using conditional classification error rate, the semantic gap between two classes, say i and j, is not symmetric, i.e. $R_{i,j} \neq R_{j,i}$. This is due to the fact that the error rate approximated using $R_{i,j}$ depends on the shape of the model of the j-th class.

Thirdly, considering that the goal of developing annotation techniques is to enable semantic-based retrieval and browsing, it is worthwhile to study to what extent the annotation framework affects retrieval and how effective it is compared with retrieving the images annotated using other methods. For this evaluation, we only study the performance with respect to those foreground concepts because the employed database is intended to provide users with images relevant to some type of animal. Since each foreground only appears once in an image, if there is any, we do not have a suitable distance function to measure the degree of similarity to rank the images, which is better than keyword matching. Therefore, we consider the following retrieval method. Given a keyword representing a foreground concept, all of the images annotated with the same keyword will be considered as the relevant ones. This results in the fact that we can not test the ranking performance of the retrieval approach, and hence can not obtain the precision and recall curve. For category ω, the precision and recall, defined as $\mathcal{P}_\omega = N_{C,\omega}/N_{R,\omega}$ and $\mathcal{R}_\omega = N_{C,\omega}/N_{G,\omega}$, are employed to evaluate the performance, where $N_{C,\omega}$ denotes the numbers of images correctly annotated by the system and thus relevant to the query ω, $N_{G,\omega}$ is the number of images belonging to the class ω according to the ground truth, and $N_{R,\omega}$ is the number of images annotated by the system with

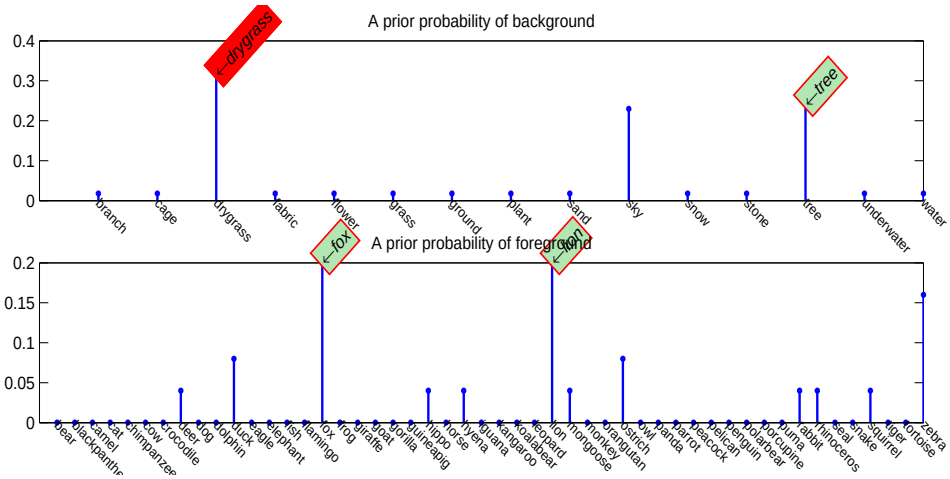
the concept ω, regardless of being correct or wrong. Finally, the average precision and recall, defined as $\mathcal{P}_{avg} = \frac{1}{W}\sum_{\omega=1}^{W}\mathcal{P}_{\omega}$ and $\mathcal{R}_{avg} = \frac{1}{W}\sum_{\omega=1}^{W}\mathcal{R}_{\omega}$, are compared among all approaches considered in our simulation.

Numerical Results

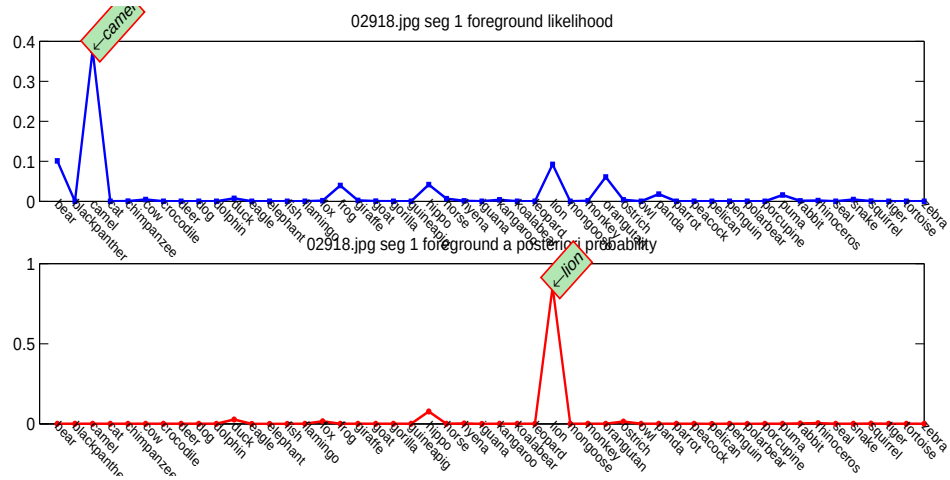
To demonstrate the advantage of the Bayesian framework over others, we compare in total three cases, including content-based annotation using SVMs (SVMA), context-based annotation (CTXA), and CBIA. Since the two-stage image segmentation brings about the availability of the information showing whether a segment contains a foreground or background concept, we consider two types of classification/annotation for each of the above three approaches to justify the improvement resulting from the two-stage image segmentation. The first type does not utilize the information obtained via the two-stage image segmentation and hence the state of nature of a to-be-classified sample may be any of the concepts in the vocabulary shown in Table. 3.1. This is referred to as all classification and ALL for short. On the other hand, the second type takes advantage of the information so that only foreground/background concepts are considered when annotating a foreground/background segment. We referred to this as separate classification and SEP for short. Therefore, six approaches are compared with each other in our study, i.e. SVMA ALL, SVMA SEP, CTXA ALL, CTXA SEP, CBIA ALL, CBIA SEP. In terms of the $P_{avg}(k)$, we also take CMRM into account in the comparative study.

Before discussing the annotation performance evaluated using the afore-mentioned criteria, the results shown in Fig. 3.5 and Fig. 3.6 are used to illustrate the underlying rationale of the proposed framework. For both examples, the results are obtained using the CBIA SEP approach because this is the most comprehensive one among all the considered ones. The first example indicates a situation in which the contextual infor-

(a) The segmentation and ground truth.



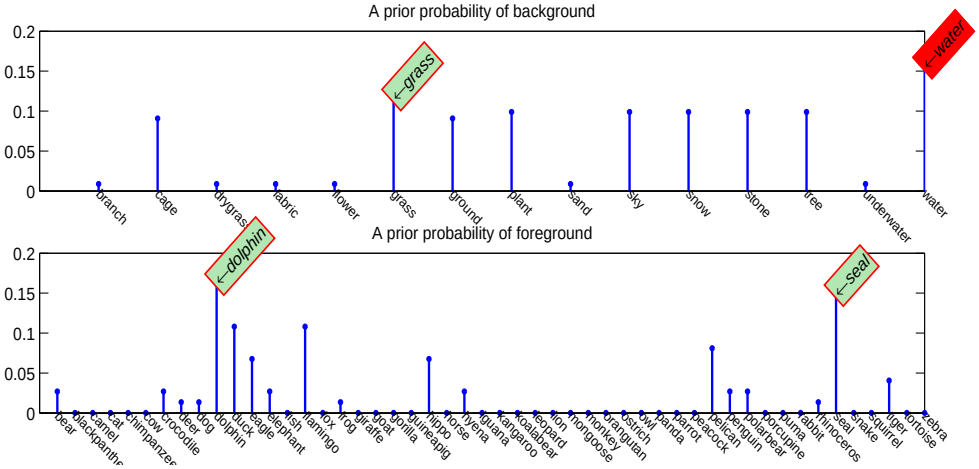
(b) The *a priori* probabilities of the image segments.



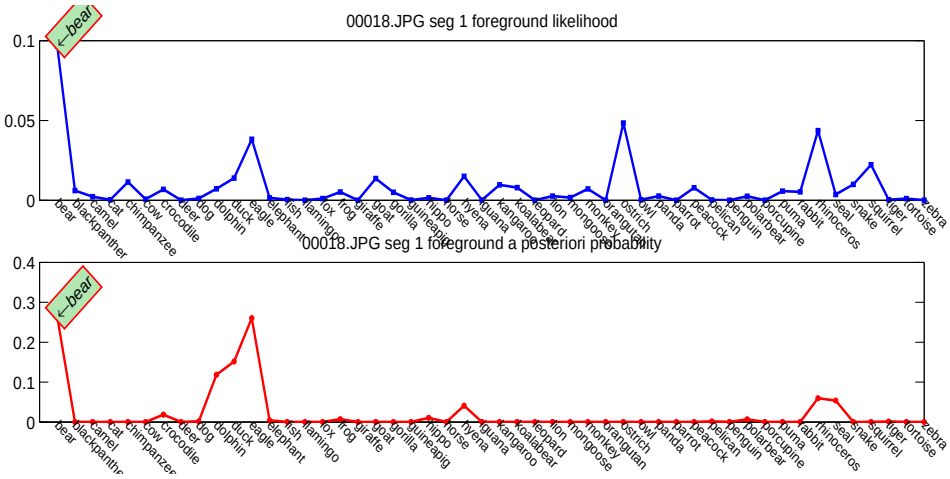
(c) The likelihood and *a posteriori* probabilities of the image segments.

Figure 3.5: Illustration of the rationale of the Bayesian framework, in which contextual information helps correct the content information.

(a) The segmentation and ground truth.



(b) The *a priori* probabilities of the image segments.



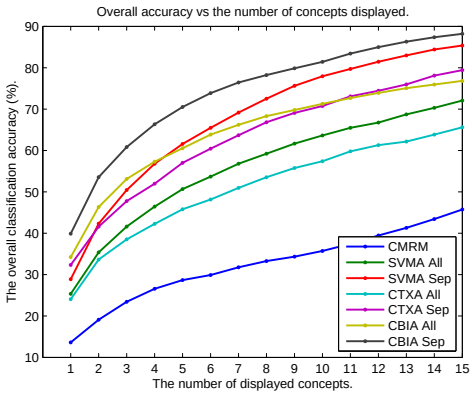
(c) The likelihood and *a posteriori* probabilities of the image segments.

Figure 3.6: Illustration of the rationale of the Bayesian framework, in which contextual information helps correct the content information.

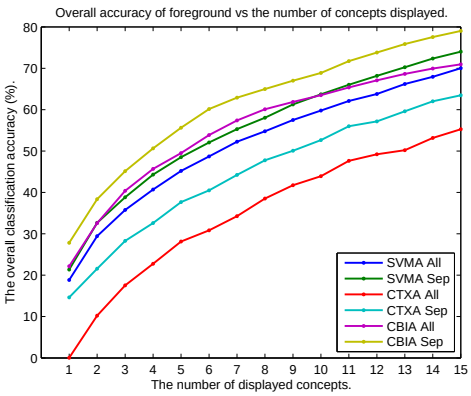
mation corrects the content information, using the example of annotating an image of a lion. Shown in Fig. 3.5(a) is the segmentation results. The *a priori* probabilities $P(\omega|\boldsymbol{I})$ of the foreground and background concepts calculated based on the context model are separately shown in Fig. 3.5(b). The likelihood $P(x|\omega)$ of the foreground segment with respect to each foreground class ω and the *a posteriori* probability $P(\omega|x, \boldsymbol{I})$ of each of them are displayed in Fig. 3.5(c). It can be observed based on the curve of likelihood that the camel is recognized as the concept assigned to the foreground segment by SVM. With the information captured by the context model, however, this value is down-weighted, while the value corresponding to lion is raised. Therefore, the Bayesian framework selects the lion as the concept for the foreground segment because its *a posteriori* probability is the highest. In contrast to the first example, the second one illustrates a case in which the content information corrects the contextual information using the annotation of an image of a bear. Given water as the background concept, *a priori* information naturally results in higher probabilities of the aquatic animals or amphibians listed in Table 3.1, such as dolphin and seal[1], which leads to a wrong decision in this example. However, the observation on the visual property of the segment refines the *a priori* knowledge in a way such that the *a posteriori* probability of bear becomes the highest.

Shown in Fig. 3.7 is the comparison among the six approaches in terms of $P_{avg}(k)$. Based on the observation that the sizes of training and testing sets for foreground and background are considerably different, as shown in Fig. 3.4(a), and the intended usage of the image database, not only do we evaluate the overall performance, we also study the performance of annotating foreground and background individually. It can be observed that, as the most comprehensive framework in question, the CBIA SEP outperforms all others regardless of how many concepts on the ranked list are examined by users, which

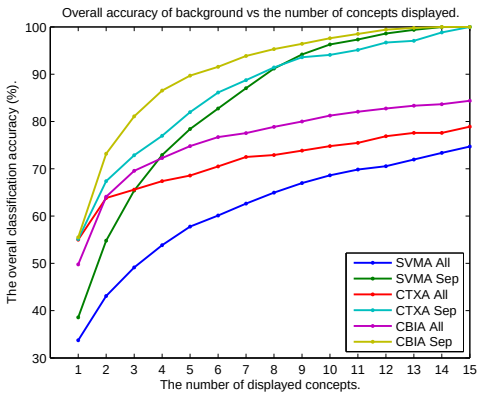
[1]We define underwater as the keyword to describe the living environment of fish.



(a) All concepts.



(b) Foreground concepts.



(c) Background concepts.

Figure 3.7: The performance evaluated using the average classification accuracy.

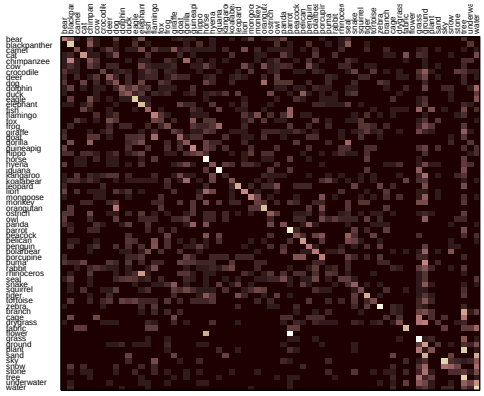
includes the case of machine-based decision making, i.e. $k = 1$. The overall performance comparison also shows that, when $k \geq 5$, SVMA SEP exhibits better performance than CBIA ALL. With respect to this observation, a closer inspection on the separate evaluation on the foreground and background shown in Fig. 3.7(b) and Fig. 3.7(c) reveals the reason. It can be seen from Fig. 3.7(c) that the performance of SVMA SEP increases much faster than CBIA ALL as the examined portion of the ranked list of concepts becomes larger, whereas this is not the case in 3.7(b). This observation itself can further be explained as follows. Based on the SEP type classification, the information of being

a background segment obtained using the two-stage image segmentation reduces the size of the set of semantic labels from 65 to 15, which almost accounts for 77% of the concept vocabulary. On the other hand, if a segment contains a foreground object, the semantic label set only shrinks by 23%. Among the remaining 50 foreground concepts, semantic gap is still very severe, whereas the gap within the 15 background concept is not. However, the above explanation only makes sense when the final decision is left to users. As long as machine-based decision is employed, the proposed CBIA framework still has best performance, for both ALL and SEP type classification. In general, the comparison indicates that the CBIA framework compensates for the drawbacks of both content-based and context-based methods, especially when the two-stage image segmentation is employed to enable SEP type classification. Moreover, the performance of annotation using CMRM is also considered in the comparison, which also indicates better performance of the proposed framework.

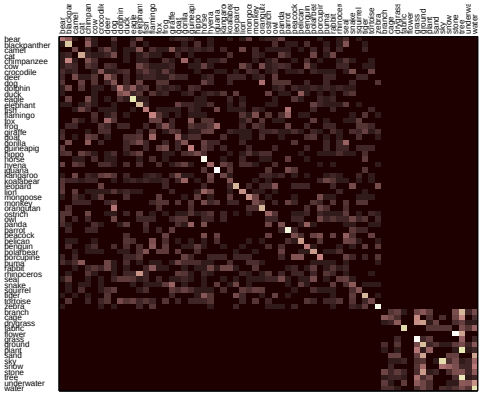
Due to the size of the concept vocabulary, we can only show the confusion matrices resulting from the six approaches and compare them using graphical illustration. In Fig. 3.8, the matrices are rendered in the way such that the brighter a block is the higher the value of the element is, at the position corresponding to that of the block. A characteristic result identifiable based on the comparison among Fig. 3.8(a) through Fig. 3.8(f) is that there are more bright blocks on the diagonal of the matrix of CBIA SEP than on those of the other approaches, which indicates higher classification accuracy resulting from this framework. In particular, the performance of annotating the background segments is improved considerably, which is consistent with the comparison shown in Fig. 3.7(c). It should be noted that the confusion matrices show more detailed information on the results corresponding to the case of $k = 1$. Moreover, the entropy values of the classification results of each semantic category using different methods are compared in Fig. 3.9.

Although the entropy values of context-based methods are lower than that of other approaches, it does not means context-based methods are better because entropy does not inform us to which semantic category the samples of the actual category are classified. Having a large subset of the samples classified into a wrong category can also lead to very low entropy. Along with Fig. 3.8, the meaningful information observed from Fig. 3.9 is that, for most of the concepts, the CBIA framework can bridge the semantic gap to some extent compared with the content-based methods, i.e. it assigns the samples to a smaller subset of the entire set of semantic classes.

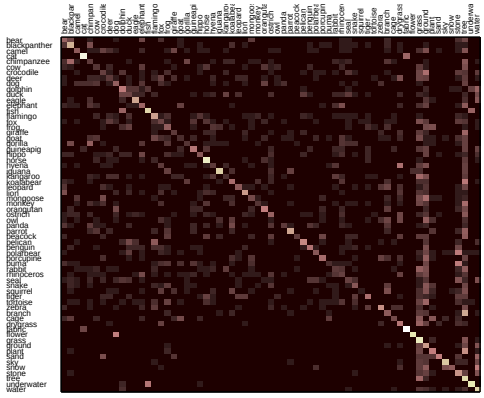
The comparison in terms of precision and recall is shown in Fig. 3.10. It is worth mentioning that the difference in terms of the recall values stems from the distinct abilities of various approaches to successfully annotate the foreground segments of the images. At the first glance, it seems that the performance of SVMA SEP is better than that of CBIA ALL, which is inconsistent with the analysis based on the $P_{avg}(k)$, where $k = 1$. In fact, although each image has one and only one foreground segment resulting in the fact that the number of foreground segments is the same as the number of images, the above conclusion is not true, which can be explained as follows. The average recall $\mathcal{R}_{avg}$ is defined as the arithmetic average of the class-specific recall $\mathcal{R}_{\omega}$, whereas the average classification accuracy P_{avg} is defined as the expected mean of the class-specific accuracy. In other words, the evaluation of the average recall does not involve any information on the sample distribution over the semantic classes, which, if considered, will give rise to another conclusion that the $\mathcal{R}_{avg}$'s of the SVMA SEP and CBIA ALL are respectively 21.37% and 22.17%. However, the comparison shown in Fig. 3.10(b) still demonstrates effectiveness and advantage of the proposed Bayesian framework in terms of the way it affects the retrieval. When it comes to the comparison based on precision, retrieval based on the annotation using ALL type classification exhibits better perfor-



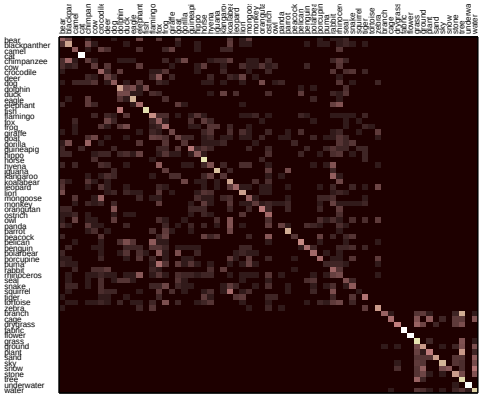
(a) SVMA ALL.



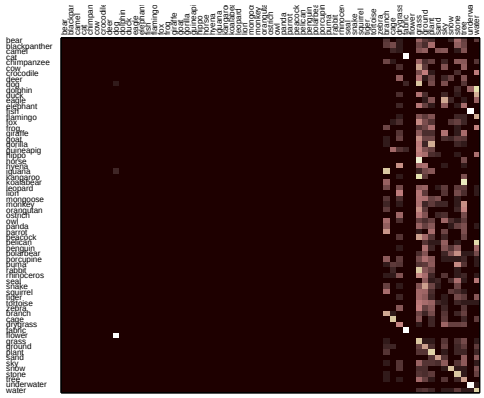
(b) SVMA SEP.



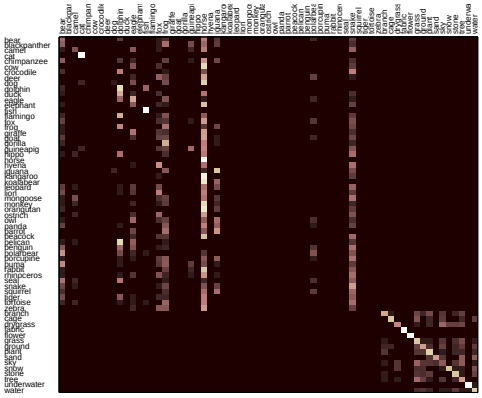
(c) CBIA ALL.



(d) CBIA SEP.



(e) CTXA ALL.



(f) CTXA SEP.

Figure 3.8: The performance evaluated using the confusion matrix.

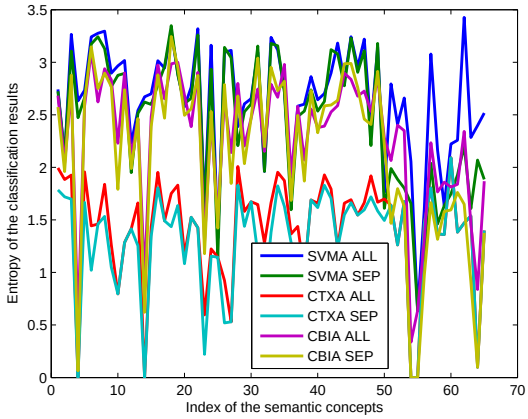
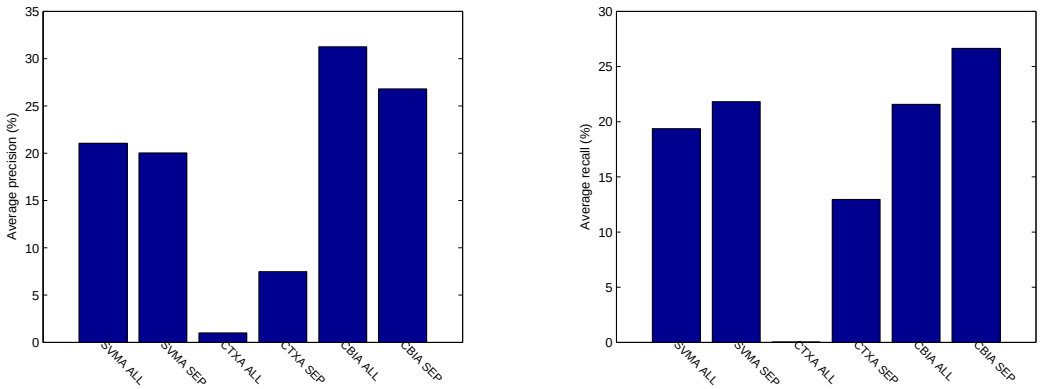


Figure 3.9: The entropy of the classification results.

mance, which is somewhat contradictory to the way it affects the annotation. Since both of the class-specific precision and recall are defined based on $N_{C,\omega}$, if the $\mathcal{P}_{avg}$ related to ALL classification is higher than that related to SEP classification and the relation in terms of $\mathcal{R}_{avg}$ is reversed, then it is very likely that $N_{R,\omega}$ related to ALL classification is lower than that related to SEP classification. We evaluate the ratio of $N_{R,\omega}$ using ALL classification to $N_{R,\omega}$ using SEP classification and the result shown in Fig. 3.11 verifies our analysis. Considering the observation one step further, we find that it can be explained as follows. When ALL classification is employed, there are in total 65 semantic categories into which the 2855 images/foreground segments are classified. However, there are only 50 classes if SEP classification is used but the total number of images is unchanged. Therefore, on average, the number of images classified into each class using ALL classification tends to be lower than that of SEP classification. In other words, in the case of ALL classification, there are many foreground segments that are classified into the semantic classes corresponding to background concepts. Nonetheless, it is worth noting that the comparison based on precision and recall justifies the advantage of the



(a) Comparison in terms of precision.

(b) Comparison in terms of recall.

Figure 3.10: The performance evaluated using precision and recall. It is the performance of a simple retrieval approach based on keyword matching. Also note that the recall of CTXA ALL is not zero but a very small value; otherwise the precision would be zero as well.

proposed CBIA framework over the content- and context-based approaches, especially when the most comprehensive framework CBIA SEP is considered.

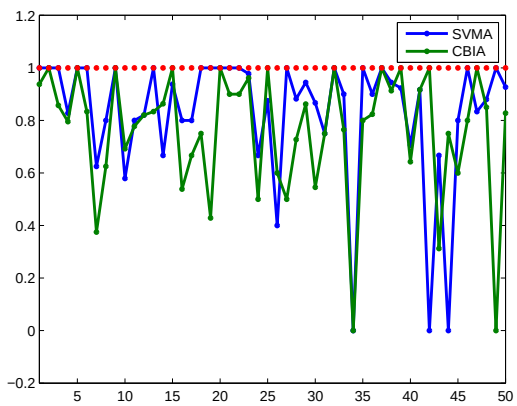


Figure 3.11: The ratio of $N_{R,\omega}$ using ALL classification to $N_{R,\omega}$ using SEP classification.

3.3.4 Summary

Driven by the technical value of effective and efficient image annotation in terms of facilitating semantic-based image retrieval and browsing, we develop an image annotation framework, which takes into account both content and contextual information. The two sources of information are integrated based on the Bayes's theorem. The low-level features of different semantic classes are represented using a set of SVMs obtained through supervised training. The contextual information is obtained using the maximum entropy estimation of a statistical model based on a set of annotated images. Appropriate modifications of the above content and context models are designed in a way such that the output of the models can be used as the likelihood and *a priori* probability, which are the fundamental components of a Bayesian framework. Experimental results based on a database featuring a variety of animals demonstrate the effectiveness of the proposed framework and its advantages over the content- and context-based approaches. In addition, the two-stage image segmentation further boosts the performance by reducing the size of the set of potential semantic classes. The proposed framework is evaluated and compared with several other approaches from many different perspective using different performance measures. The numerical results demonstrate the effectiveness of the proposed framework and its advantages over other methods. It should be noted that, in principle, the applicability of the proposed method is not limited to the database employed in our experiments, which may be considered as part of the future work.

3.4 Image Annotation Integrating Content, Context and Search

3.4.1 Overview

To remove the UMI during the annotation process, image search is incorporated into the context component of the CBIA framework. Still, the goal is to tackle the problem of region-based annotation with keyword corresponding to objects (the third task mentioned above) by integrating content-based search and machine-based visual recognition into a unified framework, where visual recognition is formulated as a Bayesian classification problem utilizing both visual information and contextual information derived from the probabilistic co-occurrence between object categories. The contribution of this work is the novel method for incorporating content-based search into the context-aware annotation process, featuring simple yet effective learning and inference. It utilizes a weighted keyword ranking to seek the most representative keywords characterizing the semantics of a to-be-annotated image. These keywords are used to refine the distribution over the object categories. The contextual information and the visual properties are utilized for Bayesian classification.

As shown in Fig. 3.12, the framework is composed of four components. First, there is a component for automatic representative keywords selection, which in turn consists of a nearest neighbor CBIR (NN-CBIR) module and a keyword ranking module. It should be noted that it is the to-be-annotated image that is used as the query of the NN-CBIR within this module. Second, a content-based module is used to evaluate the relevance of a semantic concept to the segments of the to-be-annotated image based on their own low-level visual features. Third, a context-based module evaluates the relevance of a

semantic concept to the same image based on the statistical dependence of the concept on the selected representative keywords predicted by the first component. Finally, the annotation of the image segments is considered as a Bayesian classification problem, in which the output scores of the second and third components serve as the likelihood and *a priori* probability, respectively. The details of individual components are elaborated in the following sections.

To evaluated the performance of the proposed framework, experiments are conducted using two databases, i.e. a database featuring 50 kind of animals, downloaded from Flickr™, and the widely used benchmark database for object recognition released by Microsoft™ [97]. Comparative study was conducted among the visual content-based annotation, the context-based annotation, and the annotation based on both. Experimental results demonstrated the effectiveness of the proposed framework and its advantage over others based on the annotation accuracy, the annotation precision, as well as the confusion matrix.

3.4.2 The Representative Keyword Selection Component

The essential reason for incorporating a keyword selection component is to produce some information which is needed by the context-based module, which infers the probabilities of the semantic concepts conditional on the selected ones. To focus on the discussion on the keyword selection, we defer the description on the context-based module until the next section. The proposed approach to the keyword selection is based on the following simple observation. Recall that a typical search result of CBIR using global low-level visual features usually include both relevant and irrelevant results, in which the semantic content of an irrelevant image is different from not only that of the relevant images (including

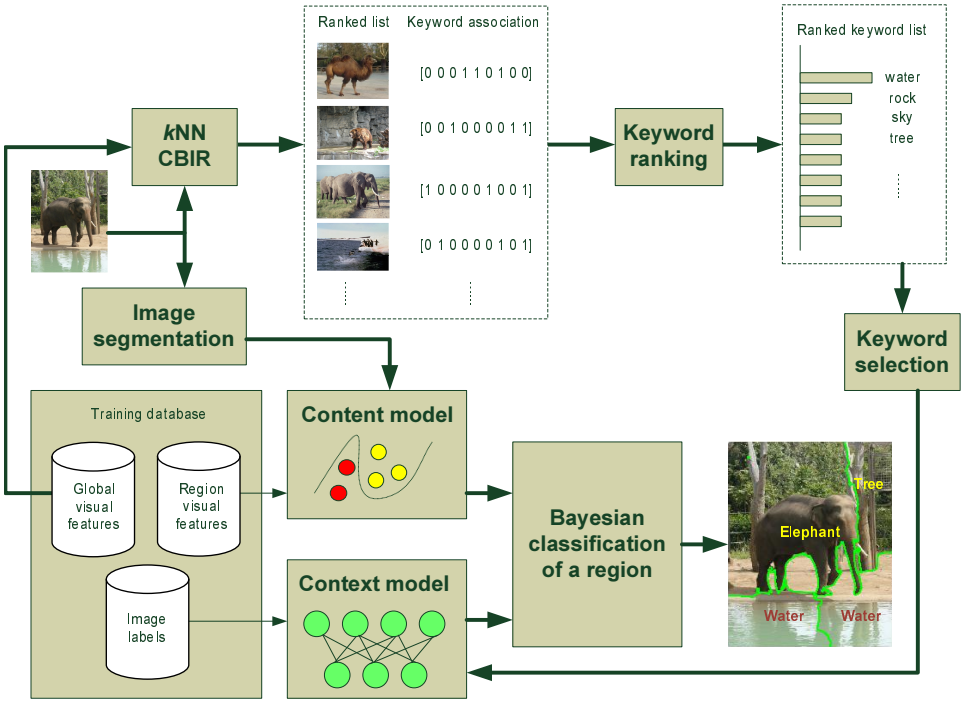


Figure 3.12: The block diagram of the proposed system. The thick arrows show the procedure of the annotation of a new image, whereas the thin arrows illustrate the training process of the framework.

the query) but also that of the other irrelevant images. This observation suggests that, if the keywords associated with all the retrieved images are pooled together, those that are most relevant to the query will stand out based on their frequency of occurrence. In other words, those irrelevant images make contribution to different sets of keywords whereas the relevant images' contribution in terms of the semantic meaning is relatively focused. Based on the above consideration, a keyword selection approach is proposed, which includes two steps, i.e. NN-CBIR using global low-level visual features followed by weighted keyword ranking. The proposed strategy for keyword ranking weights the keywords according to the dissimilarity measure between the query and the images they are respectively associated with, resulting in a more comprehensive evaluation.

Denote a keyword as ω_i, where $\omega_i \in W$ and $W = \{\omega_0, \omega_2, \ldots, \omega_{M-1}\}$ is a vocabulary with M concepts. For each of the labeled images in the training set, denoted as $I_t, t \in \{0, 1, \ldots, T-1\}$, a keyword association function $f_t(\omega_i)$ can be defined such that it takes on the value of 1 if image I_t is labeled with ω_i and 0 otherwise. After the step of NN-CBIR, the top $\tilde{T}$ images on the ranked list are used for ranking the keywords. To this end, a relevance score for each ω_i, denoted as $R(\omega_i)$, is calculated by

$$R(\omega_i) = \sum_{t=0}^{\tilde{T}-1} \exp(-\frac{d(I_t, Q)}{r}) f_t(\omega_i), \tag{3.11}$$

where Q represents a query image, $d(I_t, Q)$ is the dissimilarity measure between I_t and Q, which is L1-Norm in the presented work, and r is a parameter that can be adjusted for a dataset containing a certain set of semantic concepts. By the weighted keyword ranking, the M keywords end up being in such an order that their relevance scores satisfy $R(\omega_{i_0}) \geq R(\omega_{i_1}) \geq \ldots \geq R(\omega_{i_{M-1}})$, where $i_k \in \{0, 1, \ldots, M-1\}$ is the index of the k-th concept on the ranked list. In our present study, only the top 2 concepts are considered during the subsequent steps of the annotation process and an adaptive determination of the cut-off number is left as our future study.

The basic NN-CBIR is employed based on the following consideration. First, as retrieval is only used as one step of the annotation process, it is expected to be efficient from the practical point of view. Second, although more sophisticated (dis)similarity measure can improve the retrieval performance and in turn the automatic keyword selection, research on object recognition and scene understanding based on large scale database [98] showed the effectiveness of the k nearest neighbors methods. Last but not least, the keyword selection only relies on weakly labeled training images and nowadays the acquisition of a considerably large amout of such images is feasible due to the popularity of

the online photo sharing web sites, e.g. Flickr™.

3.4.3 The Visual Content and Context Analysis

The functionality of this component is relatively independent of that of the keyword selection component. Its major task is to produce a score which characterizes the relevance of a semantic concept to an input feature vector, which is extracted from an image segment. Therefore, the content-based module consists of a set of models, each of which corresponds to a certain semantic concept in the vocabulary. SVMs are employed as described in the section of content model of the Bayesian framework. As a result, each of these SVMs is used to calculate the distance between the testing sample and the decision hyperplane of of the SVM, followed by the conversion of this distance to the value of a PDF expressed in (3.4). The context-based component utilizes the statistical dependence between different semantic concepts to infer the relevance of one semantic concept given another. To be specific, the objective of using this component is to calculate the conditional probability $P(\omega_i|\omega_s)$, where $\omega_s \in W$ is a set of keywords selected by the keyword selection component and can be considered as the background information $\boldsymbol{I}$ of the Bayesian framework. Hence, replacing $\boldsymbol{I}$ with ω_s in (3.6), we obtain

$$P(\omega_i|\omega_s) = \frac{P(Y_{\omega_i} = 1|Y_s = 1)}{\sum_{j=0}^{M-1} P(Y_{\omega_j} = 1|Y_s = 1)}. \tag{3.12}$$

Finally, the content and context are integrated through (3.3) and the decision is made according to (3.10).

3.4.4 Experiments

Databases, Performance Measure, and Experimental Setup

To evaluate the performance of the proposed framework, we employed two databases, which are described as follows.

- **Animal5k.** This database consists of 5000 images collected from the Internet, featuring 50 categories of animals, with each category containing 100 images. A semantic concept vocabulary of 65 concepts are defined, including those corresponding to the 50 kinds of animals as well as those for the background in common natural scene. From the 100 images of each animal category, 60% are randomly selected as the training images, and the rest are testing images. The images are segmented using normalized cut [55].

- **MSRC.** It is a benchmark database for object recognition released by the Microsoft Research Cambridge [97]. It includes 591 images and 23 semantic concepts. Since this database is relatively small, only 50% of the images containing each concept are randomly chosen as training images and the rest are used as testing images.

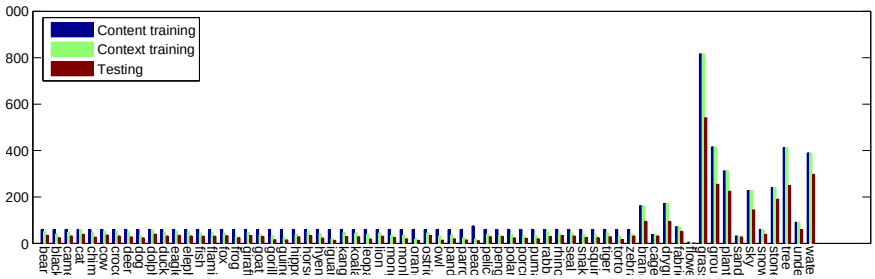
More details on the image usage of the two databases are given in Table. 3.3 and Fig. 3.13. In addition, the co-occurrence patterns of the training data are shown in Fig. 3.14. For the content-based search, the 102-dimensional global image features include color histogram, color layout, Fourier descriptors, and Gabor wavelets. For the visual content component, region features include color moment and texture moment, resulting in a 26-dimensional vector.

We employed annotation accuracy and average precision to evaluate the effectiveness of the proposed framework. The former is defined as the ratio of the number of correctly

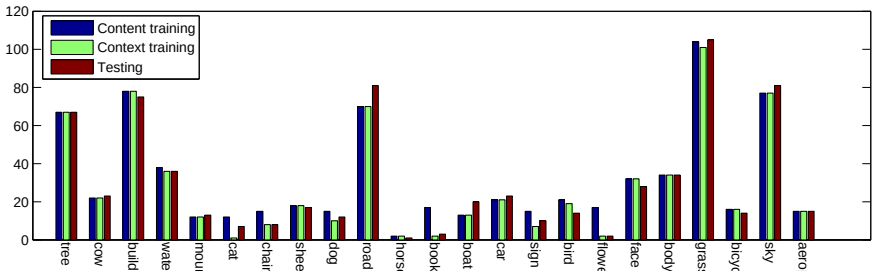
Table 3.3: Image usage of the two databases.

Data set	Animal5k	MSRC
Training set for content model	3000	296
Training set for context model	2046	228
Testing set	1343	233

classified image segments to the total number of image segments in the testing set. For the latter, the precision of classifying the segments of each concept category is evaluated followed by taking the arithmetic average, where the precision is defined as the ratio of the number of correctly classified image segments of a certain concept to the total number of segments classified into that category. As shown in Fig. 3.13, the distributions of the testing samples of both databases exhibits serious imbalance. This directly results in the fact that the performance measurements of different concept classes have different confidence levels. Therefore, the annotation accuracy rather than average recall is chosen. In fact, the theoretical formulation of the average classification error which is minimized by the maximum *a posteriori* probability (MAP) rule is the class-specific error rate averaged based on the distribution of the classes. The annotation accuracy is based on exactly the same idea, although the class distribution is empirical. To further inspect the performance on each specific semantic concept, we also consider the confusion matrix, in which the diagonal elements are the concept-specific recall values. In total, three annotation approaches are compared in our experimental study, which are content-based annotation (CTNA), context-based annotation (CTXA), and the proposed Bayesian annotation (BA). For each annotation approach, two scenarios are considered. In the first one, no object of interest is specified by a user and hence all the segments are treated in the same way. We refer to this as ALL for short. The second scenario



(a) Animal5k



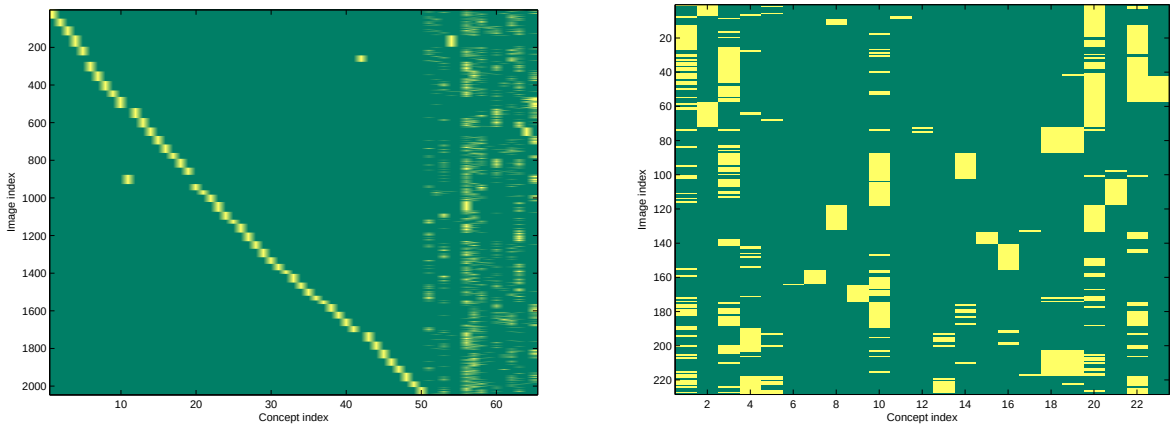
(b) MSRC

Figure 3.13: The segment distribution over classes of the two databases.

assumes that a user selects an object of interested, which will be annotated using one of the keywords belonging to the subset of semantic concepts characterizing foreground objects. This is referred to as SEP for short.

Experimental Results

Shown in Fig. 3.15(a) and Fig. 3.15(b) are the annotation accuracy and average annotation precision. It can be observed that for both databases and both scenarios the proposed Bayesian approaches outperform both of the content-based approach and the context-based approach. To save space, the results for both scenarios are shown in the same figure for each database; however, it should be noted that the performance should be compared among different approaches within each individual scenario. There is not as much improvement on the MSRC database as there is on the Animal5K database.



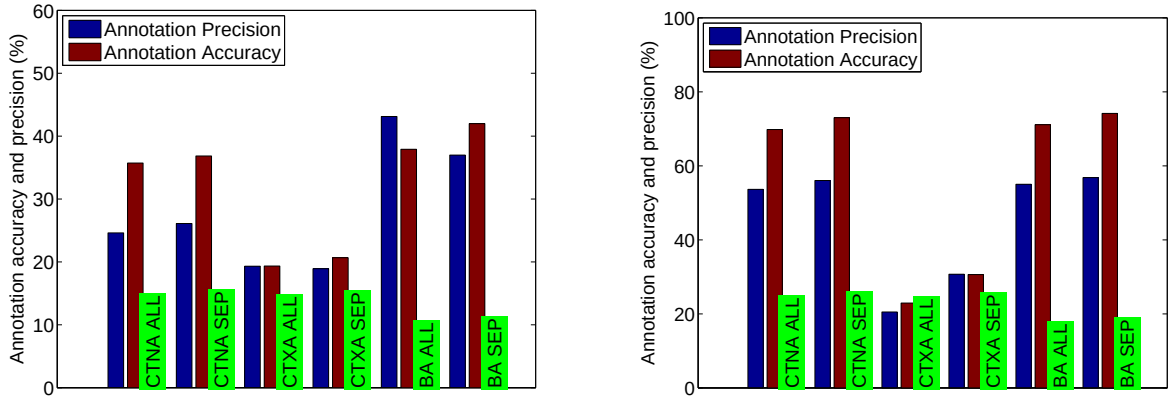
(a) Animal5k (b) MSRC

Figure 3.14: The co-occurrence patter of the training data.

The reason is that so far only feature functions defined over a single variable have been considered. According to the comparison based on the confusion matrix shown in Fig. 3.16, it can be seen that, using the Bayesian approaches, there are more elements on the diagonals, of which the colors are red or closer to red, indicating the better performance with respect to those classes in terms of recall. To visually inspect the annotation results, we selected a few examples from those of the Animal5k database, which are shown in Fig.3.17.

3.4.5 Summary

In this section, CBIR is incorporated into the CBIA framework to select salient keywords which can be used to generate the contextual information without human users being involved in the annotation process. With the representative keywords selected by the search component, the *a priori* distribution of the object categories can be calculated through the maximum entropy approach. This *a priori* information is integrated with the likelihood evaluated using visual content models adapted from SVMs through the



(a) Animal5k (b) MSRC

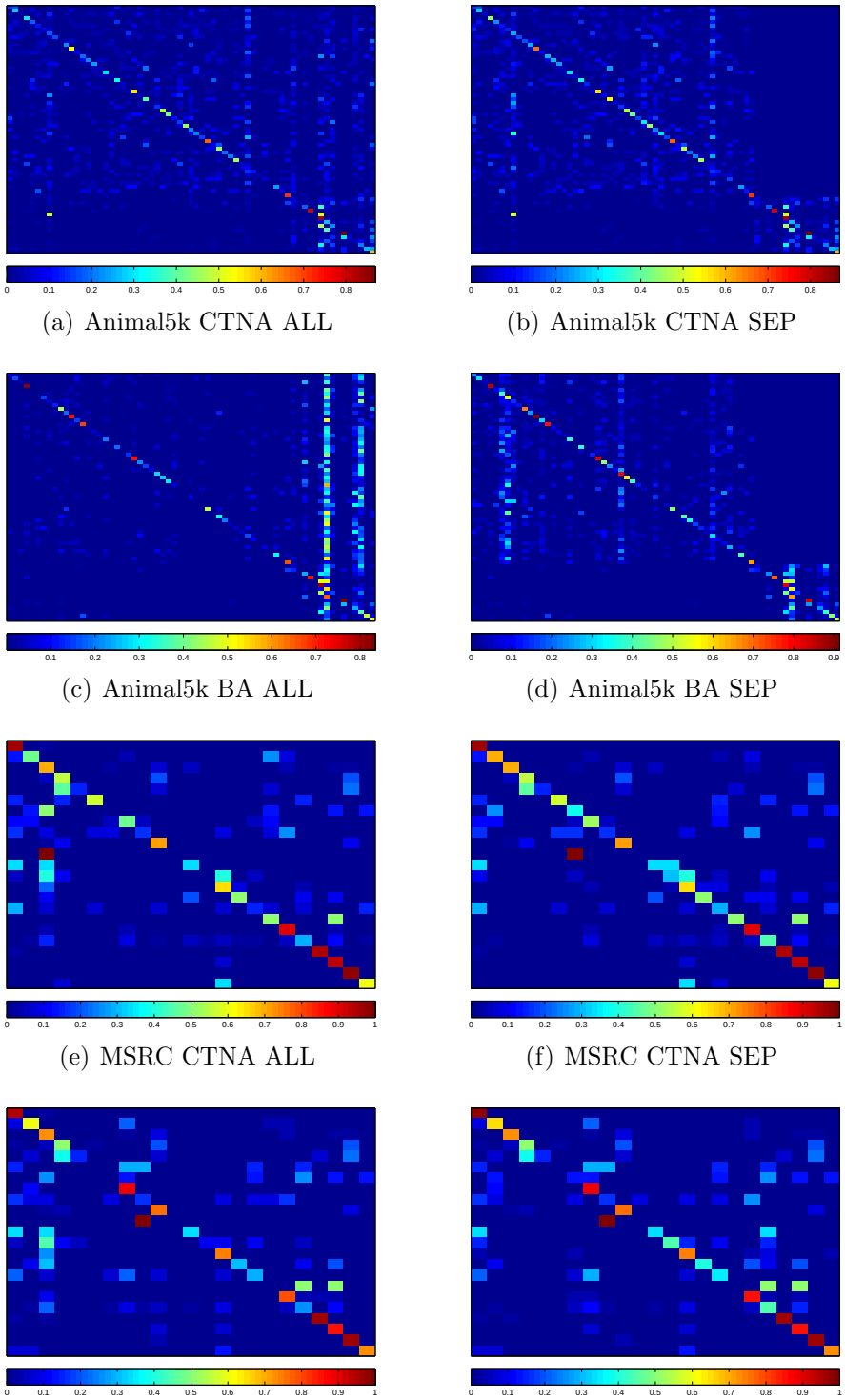
Figure 3.15: Annotation accuracy and average precision.

Bayesian classification. Experimental results demonstrated the advantages of fully automatic Bayesian framework over the visual content-based and context-based approaches.

3.5 The Application to Image Retrieval

3.5.1 Overview

The Bayesian framework is also applied to tackle the semantic gap of image retrieval by integrating short-term relevance feedback (STRF) and long-term relevance feedback (LTRF). The STRF refers to the user interaction during a retrieval session consisting of a number of feedback iterations, such as the query movement and the query feature reweighting. On the other hand, the LTRF is the estimation of a user history model from the past retrieval results approved by previous users. Experiments with the proposed framework has demonstrated that the LTRF plays a key role of refining the degree of relevance of the candidate images in a database to a query. In the proposed image retrieval framework, the STRF and LTRF play the roles of refining the likelihood and



(a) Animal5k CTNA ALL (b) Animal5k CTNA SEP

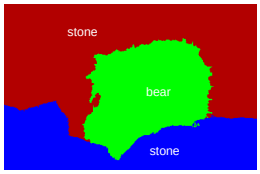
(c) Animal5k BA ALL (d) Animal5k BA SEP

(e) MSRC CTNA ALL (f) MSRC CTNA SEP

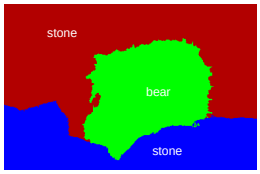
(g) MSRC BA ALL (h) MSRC BA SEP

Figure 3.16: Confusion matrices of various annotation methods.

(a) Bear

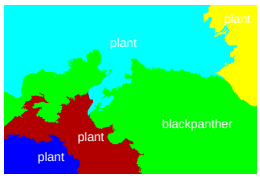


(b) CBIA SEP

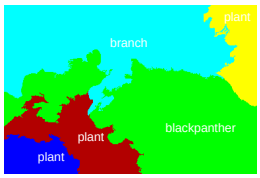


(c) Ground truth

(d) Black panther

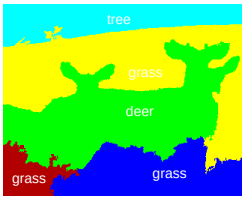


(e) CBIA SEP

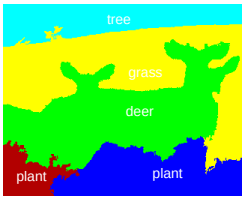


(f) Ground truth

(g) Deer

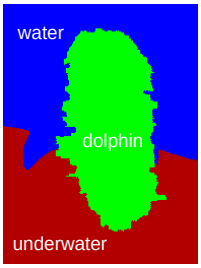


(h) CBIA SEP

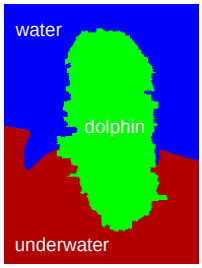


(i) Ground truth

(j) Dolphin



(k) CBIA SEP

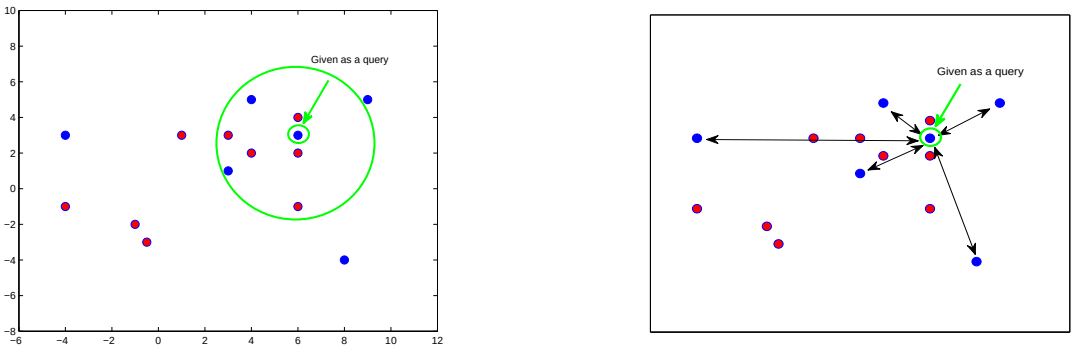


(l) Ground truth

Figure 3.17: Some examples of annotation results. For each example, the figure in the middle is the annotation results and the one on the right-hand side is the ground truth.

the *a priori* information, respectively, and the images are ranked according to the *a posteriori* probability. Since the estimation of the user history model is based on the principle of collaborative filtering, the system is referred to as a collaborative Bayesian image retrieval (CLBIR) framework. By exploiting the past retrieval results, it can be considered as a CBIR system with memory, which incrementally learn the high level knowledge provided by human users.

The underlying rationale of applying the Bayesian framework to image retrieval can be illustrated using Fig. 3.18, of which the gist is to boost the retrieval performance using some information extracted from the retrieval history (In the rest of this section, past retrieval results, retrieval history, and user data are terms that are used interchangeably.). As mentioned earlier, the two types of similarity measure are complementary to each other. Specifically, the similarity measure by the content-based component illustrated by the low-level feature space in Fig. 3.18(a) suffers from the semantic gap which can be alleviated using the contextual information. The links between relevant images in Fig. 3.18(b) are estimated by utilizing the co-occurrence of relevant images in the past retrieval results. At the same time, the contextual information can only be acquired by learning from the knowledge accumulated through the content-based component. Therefore, the goal of the proposed framework is to utilize these two types of information jointly and effectively. The CLBIR framework, illustrated in Fig. 3.19, seamlessly integrates the content-based and the context-based methods into a mathematically justifiable framework. In the beginning, there is no available retrieval history to learn the context model but the system can still work using the content-based component and incrementally accumulate the retrieval results. When past retrieval results are available, the context component of the system performs LTRF by extracting information from the data gradually, which can be considered as a knowledge accumulation process. When a



(a) Semantic gap exists in the content domain.

(b) There might not be sufficient data to extract accurate contextual information.

Figure 3.18: The similarity measure in the content and context domains.

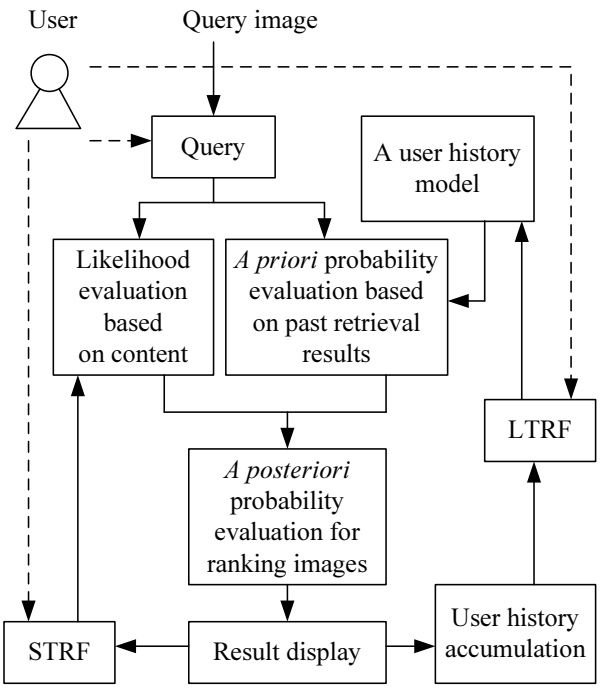


Figure 3.19: The block diagram of the CLBIR framework. The solid and dashed directed lines indicate the information flow and the human-controlled components in the framework, respectively.

user presents a query, the content component of the system learns the user's information need from the query through similarity measure and STRF. If the context component has

been trained by the time a user queries the database, the system is capable of integrating the useful information predicted using the context component and that learned using the content component. The *a posteriori* probability evaluated using the CLBIR framework is used to rank the images in the database.

3.5.2 The Content and Context Components

The NN-CBIR Content Component

As mentioned before, the goal of the content analysis is to obtain the likelihood that a certain candidate image is relevant to the query. To this end, we adopt two different approaches to CBIR, i.e. NN-CBIR based on L1-Norm and SVM. In terms of the approaches to STRF associated with these two types of content components, query point movement and active learning are employed, respectively. Hence, the second is referred to as SVMAL-CBIR in what follows for short.

The mechanism of NN-CBIR is to return the top K images on the list, which is ranked based on the similarity measure between the feature of the query and that of each of the candidate images, where $K \ll N$. In our framework, the L1-Norm is used as a distance function, which is defined as

$$d(\boldsymbol{x}_q, \boldsymbol{x}_\omega) = |\boldsymbol{x}_q - \boldsymbol{x}_\omega| = \sum_{j=1}^{d} |x_{q,j} - x_{\omega,j}|, \tag{3.13}$$

where $\boldsymbol{x}_q$ and $\boldsymbol{x}_\omega$ denote the desciptor vector of a query image and a candidate image. The likelihood of the query image with respect to each of the candidate images can be evaluated by substituting (3.13) into (3.4). For STRF, the refined query based on query

point movement can be expressed as

$$\boldsymbol{x}_q^t = \alpha \boldsymbol{x}_q^{t-1} + \beta \left(\frac{1}{N_P} \sum_{u=1}^{N_P} \boldsymbol{x}_u^{t-1} \right) - \gamma \left(\frac{1}{N_N} \sum_{v=1}^{N_N} \boldsymbol{x}_v^{t-1} \right), \tag{3.14}$$

where α, β, and γ are pre-selected parameters, and N_P and N_N are the numbers of positive and negative examples within the retrieved set of images after the $(t-1)$th iteration, $t \geq 2$, and the superscript of the feature vectors indicates the number of retrieval iterations.

The SVMAL Content Component

In order to demonstrate the flexibility of the CLBIR framework to use different types of similarity measure for the content component, SVM is selected, in addition to the nearest neighbor method, due to its theoretical and practical value. Theoretically, a learning machine, defined as a set of parameterized functions and trained based on the maximal margin principle results in a decision function which can be expressed as the linear combination of the support vectors, which are the training samples closest to the decision hyperplane. By maximizing the margin, the trained SVMs minimize the generalization error, while maintaining the minimum empirical error. In addition, due to the linear non-separability of most of the practical problems, the advantage of SVMs lies with the transformation from a low-dimension space to an arbitrarily high-dimensional space introduced by a properly selected kernel function. A comprehensive tutorial on SVM can be found in [99].

Given a set of training samples, denoted as $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_T, y_T)\}$, where $\boldsymbol{x}_i \in \mathbb{R}^d$, $y_i \in \{-1, 1\}$ is the ground-truth label of $\boldsymbol{x}_i$, and $i \in \{1, 2, \ldots, T\}$. The searching for the optimal hyperplane in the weight space Ψ can be accomplished by

solving either the primal optimization problem or the dual optimization problem. When the kernel function is involved, the optimal hyperplane can be represented as

$$\begin{aligned} f(\boldsymbol{x}) &= \sum_{i=1}^{T} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b \\ &= \sum_{i=1}^{T} \alpha_i y_i \Phi(\boldsymbol{x}_i)^T \Phi(\boldsymbol{x}) + b \\ &= \boldsymbol{w}^T \Phi(\boldsymbol{x}) + b \end{aligned} \tag{3.15}$$

where $K(\boldsymbol{x}_i, \boldsymbol{x})$ is the kernel function, α_i is the Lagrangian multiplier which can determined by solving the dual optimization problem, b is the bias which can be determined using the KKT complementarity condition, $\Phi(\boldsymbol{x}) : \mathbb{R}^d \mapsto \mathbb{R}^p$, $d \ll p$, is the transformation, and $\boldsymbol{w}$ is the weight vector. Since a properly selected kernel function satisfies the Mercer's condition, it is not necessary to know the specific form of $\Phi(\boldsymbol{x})$ because both the optimal hyperplane and the objective function depend only on the inner product of two feature vectors in the transformed space Ω. In our study, the radial basis function (RBF) is chosen to be the kernel function. Due to the sparse sample problem of the relevance feedback in general CBIR, the methodology of active learning was introduced into the human-machine interaction for STRF, where the most informative images are shown to request user-provided labeling, resulting in the SVMAL-CBIR [19]. These images are those reside closet to the present optimal hyperplane. The underlying idea of active learning with SVM is to choose the unlabeled samples to reduce the version space as fast as possible, where the version space is defined as the region in Ψ corresponding to the hyperplane capable of perfectly classify the training samples in Ω. This is realized by selecting, in Ω, the unlabeled samples that are closest to the current learned hyperplane, i.e. the samples with minimum $|\boldsymbol{w} \cdot \Phi(\boldsymbol{x})|$. Such samples will be included in the training

set for the next iteration of the SVM learning, which is the STRF scheme for SVMAL-CBIR. The evaluation of the likelihood of a query image with respect to a candidate image is performed through (3.4).

3.5.3 Experiments

Experimental Setup

To guarantee the diversified image content, which is a typical situation of image retrieval in a large general domain, we randomly selected 200 classes from the COREL image collection, with 50 images in each class. The resultant 10000 images and the vendor-defined categories were used as the database and the ground truth for evaluating the performance. From the database, 10 queries are selected from each of the 200 classes, resulting in 2000 queries are selected, each of which is composed of two different images. Under the query-by-example retrieval paradigm, the average of the features of the two images is used as the feature of an exemplar image. To facilitate the subsequent elaboration, the query subsets which consist of the first five queries, the sixth through the eighth, and the ninth and the tenth in each class, are denoted T_A, $T_{B,1}$, and $T_{B,2}$, where $|T_A| = 1000$, $|T_{B,1}| = 400$, and $|T_{B,2}| = 600$. Such a query set selection guarantees that the system trained using the LTRF will be tested based on previously unseen samples. T_A was used when there is no accumulated high-level knowledge, i.e. before LTRF happens. In such a case, only STRF is involved, and the NN-CLBIR and the SVMAL-CLBIR are essentially the same as the NN-CBIR and SVMAL-CBIR because the *a priori* distribution of the candidate images is uniform. After the initial LTRF, the CLBIR systems are expected to present better performance in general thanks to the accumulated knowledge, while the STRF still improves the results with respect to each specific query. $T_{B,1} \cup T_{B,2}$, in-

Table 3.4: Summarization of the feature extraction.

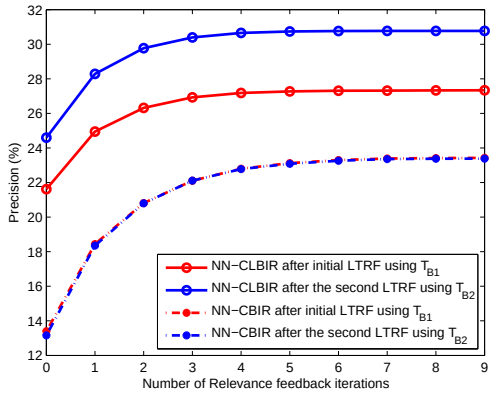
Color Feature	
Color Histogram	16, 4, 4 bins in H, S, V channels
Color Layout	An image is partitioned into 8×8 blocks, 6, 3, 3 coefficients in Y, Cb, Cr channels
Texture Feature	
Gabor Wavelet	4 scales and 6 orientations

cluding 1000 image, was used to verify the improvement after the initial LTRF. During the operation of the CLBIR systems, the new retrieval results after the initial LTRF are gradually accumulated, and a second LTRF can be carried out upon a certain point. The retrieval results corresponding to $T_{B,1}$ were used to perform an incremental update of the system, i.e. the second LTRF, after which the performance was evaluated using $T_{B,2}$.

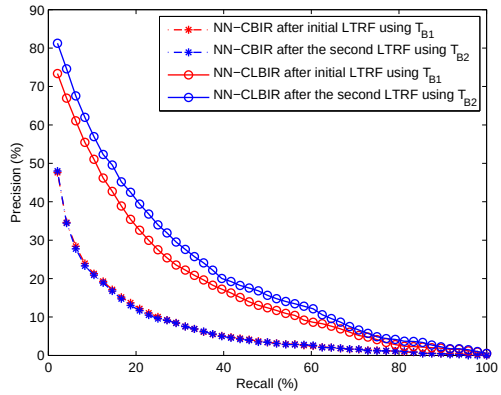
To capture various visual properties of the images, three types of low-level descriptors are selected, including global color histogram in Hue-Saturation-Value (HSV) space, color layout in YCbCr space [100], as well as Gabor wavelet [101]. The detailed information on the feature extraction is outlined in Table 4.1.

Numerical Results

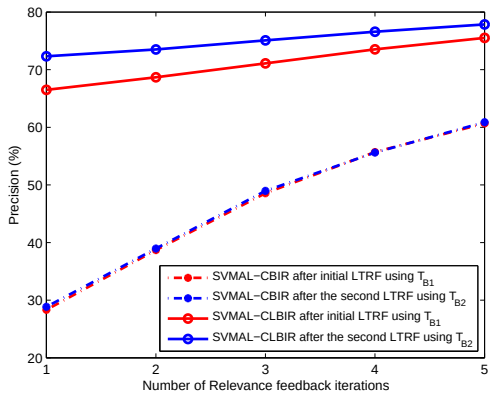
Shown in Fig. 3.20(a) is the comparison between NN-CBIR and NN-CLBIR in terms of the average precision P_{avg} as a function of the number of iterations of STRF, where the precision is defined as $P = \frac{N_C}{N_R}$, where N_C and N_R are the numbers of relevance images and retrieved images, respectively. We adopted $N_R = 48$ in this case. Using query set $T_{B,1}$, the improvement due to the LTRF based on the past retrieval results with respect to the query set T_A is obvious, and the effect of STRF can also be observed. After the second LTRF, the performance of NN-CLBIR using query set $T_{B,2}$ is further enhanced



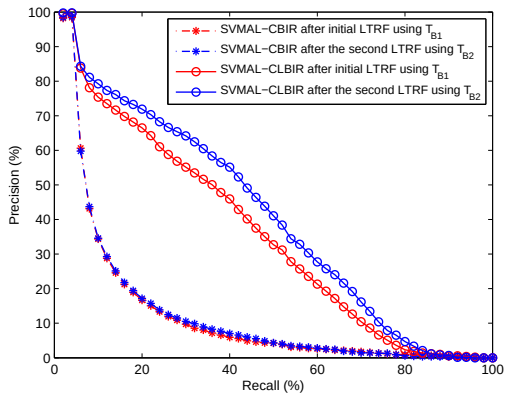
(a) Comparison between the performance of NN-CBIR and NN-CLBIR in terms of the average precision versus the number of relevance feedback iterations.



(b) Comparison between the performance of NN-CBIR and NN-CLBIR in terms of the precision versus recall after the first retrieval iteration.



(c) Comparison between the performance of SVMAL-CBIR and SVMAL-CLBIR in terms of the average precision versus the number of relevance feedback iterations.



(d) Comparison between the performance of SVMAL-CBIR and SVMAL-CLBIR in terms of the precision versus recall after the first retrieval iteration.

Figure 3.20: Objective evaluation on the performance improvement resulting from the proposed approach. a) and c) Comparison in terms of the PRC after the first retrieval iteration. b) and d) Comparison in terms of the precision as a function of the number of RF iterations.

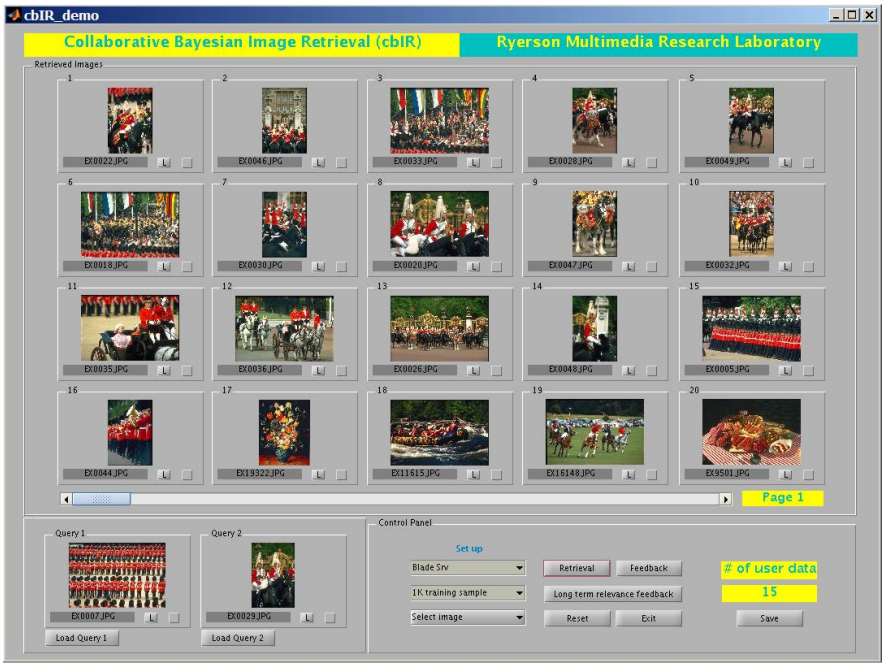
resulting from more accumulated knowledge through the LTRF. Based on the same query set, the performance of NN-CBIR remains unchanged. To test the performance in terms of ranking ability, we employed the precision-versus-recall curve (PRC), where the recall

is defined as $R = \frac{N_C}{N_G}$, where N_G is the number of images in the same semantic class as that of the query. The precision is averaged over all queries at each different recall value. The PRC after the initial retrieval was shown in Fig. 3.20(b). Higher precision value at a certain recall indicates more relevant images being ranked ahead of irrelevant ones, i.e. to reach the recall value, a smaller set of retrieved images has to be gone through. Based on this fact, the advantage of the integration of user history as high-level knowledge with the content analysis can be demonstrated based on the comparison in Fig. 3.20.

The comparison shown in Fig. 3.20(c) and Fig. 3.20(d) is for the same purpose of performance evaluation as that described above, and the difference lies with the approach to the content analysis for the likelihood computation, which is based on the output of the SVM employed for the active learning-based STRF. In this case, we adopted $N_R = 20$ for the evaluation of precision as a function of the number of STRF iteration, and $N_C = 50$ for the evaluation of PRC. Since the initial retrieval is just random ranking, the precision was evaluated starting from the first STRF iteration. Still, we can observe the improvement resulting from the integration through the Bayesian framework.

Subjective Evaluation

An interface with the NN-CLBIR enabled has been implemented to demonstrate the effectiveness of the proposed framework in terms of performance improvement by the accumulation of user history. Illustrated in Fig. 3.21(a) and Fig. 3.21(b) are the top 20 retrieved images using NN-CLBIR. Shown in the figure on the left is the result obtained using a system, whose *a priori* knowledge was extracted from 1000 user data, while on the right, the result is based on the *a priori* knowledge learned from 1400 user data. The query is selected from the semantic class of the theme soldier, and the last 4 images do not belong to this class in Fig. 3.21(a). Nonetheless, all of the top 20 images are relevant



(a) Based on the user history model trained using 2000 past retrieval results.

(b) Based on the user history model trained using 3200 past retrieval results.

Figure 3.21: Retrieval results for subjective evaluation on the performance improvement resulting from more user history.

to the query.

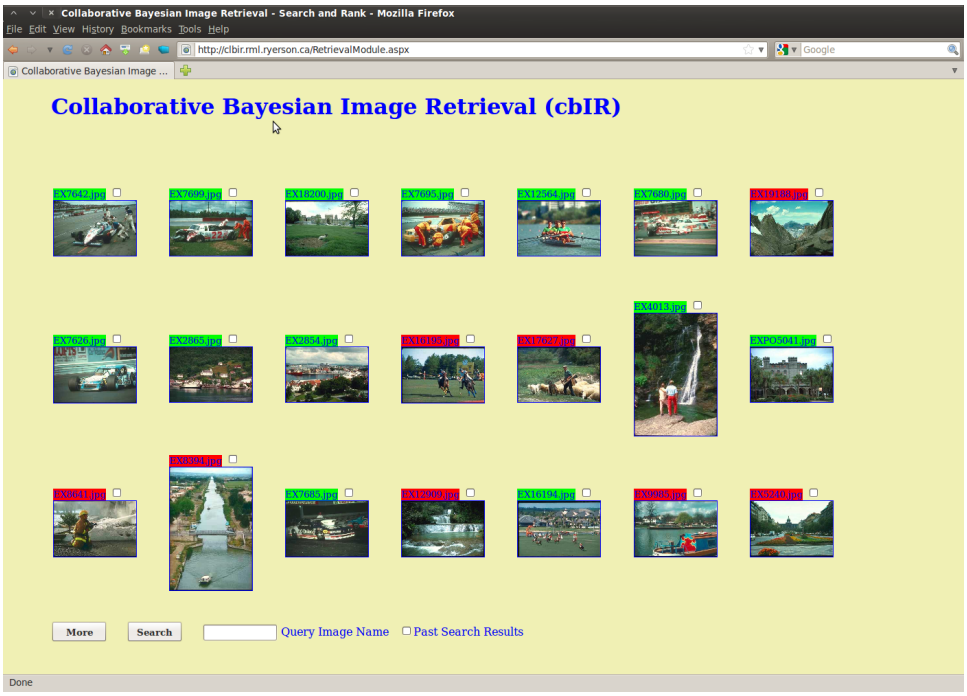
3.5.4 A Prototype System of the Search Engine based on CLBIR

Beyond the experiment-based study of the CLBIR framework, a prototype system of the CLBIR has been implemented. The image database is composed of 40000 images obtained from the Corel image collection. There are in total 400 semantic classes, each of which has 100 images. For similarity measure using the visual content component of the framework, we implemented global color histogram, color layout, Fourier descriptors, and Gabor wavelets, which are used as the low-level visual features for image representation in the feature space. The four descriptors of an image are cascaded into a single vector of 102 dimensions. L1-Norm is employed as the distance function. To enable the context component, we need past retrieval results which are obtained using the visual content component. To this end, we collected 8000 retrieval results using the content component, for which 20 query images are randomly selected from each of the 400 semantic classes. The context model is theoretically a 40000×40000 matrix. The sparse matrix representation was employed because there are many zero-valued entries due to the limited number of retrieval results. The demo was implemented as a web application using ASP.NET with C# programming language. The application is hosted by a machine with a 2.4GHz Intel Core 2 Quad CPU and 4GB RAM, and the operating system is Windows Server 2003 R2. There are 21 images on a page of the ranked list of images, organized in a layout of 3 rows by 7 columns. The file name of an image displayed on the web page is shown on top of the image. If the background color of the title of an image is green/red, it means that contextual information is available/unavailable for that image. A user can choose to use the context information or not. The URL of the

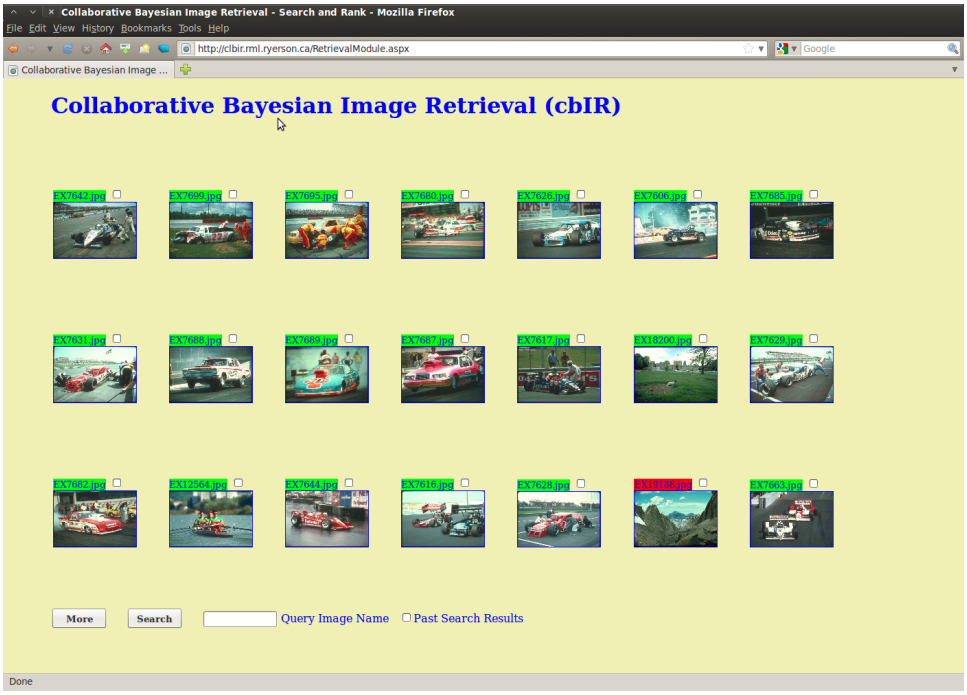
application is http://clbir.rml.ryerson.ca/main.htm. A video demo is also available at http://www.youtube.com/watch?v=SEsT9c3kzLw. An example of the search results is illustrated in Fig. 3.22.

3.5.5 Summary

The STRF and LTRF are integrated through the proposed CLBIR framework. To be specific, the content and context, obtained via STRF and LTRF, are combined through a Bayesian framework. The CLBIR framework can be considered as a CBIR system with memory, which can incrementally accumulate high-level semantic knowledge assisting in bridging the semantic gap in the future retrieval performed by prospective users. Two particular instances of the proposed framework has been implemented for experimental evaluation, which are SVMAL CLBIR and NN-CLBIR. Simulation results demonstrated the effectiveness of the combination of the content-based and content-independent information, which include the improvement resulting from learning a user history model based on more accumulated knowledge, i.e. LTRF, and that by STRF during each retrieval session. Future work will be focused on seeking a more accurate approach to estimating the user history model.



(a) Result of CBIR.



(b) Result of CLBIR.

Figure 3.22: An example of the comparison between the search results of CBIR and CLBIR using the prototype system.

Chapter 4

Image Retrieval by Integrating Audio and Visual Information

4.1 Introduction

In this chapter, we propose a new framework of multi-modal image retrieval, which utilizes the information in both audio and visual domains. By considering the audio information as a kind of holistic background context, the problem can again be tackled from the angle of the integration of content and context. Therefore, the general Bayesian framework presented in the last chapter is employed, which the contextual information is induced from the characteristic audio features of different objects, while their visual features are the input of the content analysis. A database of 4400 images featuring 50 kinds of animals is employed in our experiments. Based on comparative experimental evaluation, the numerical results demonstrate better performance resulting from the proposed fusion of audio and visual information. A guideline for further study on the framework is also discussed based on the experimental results.

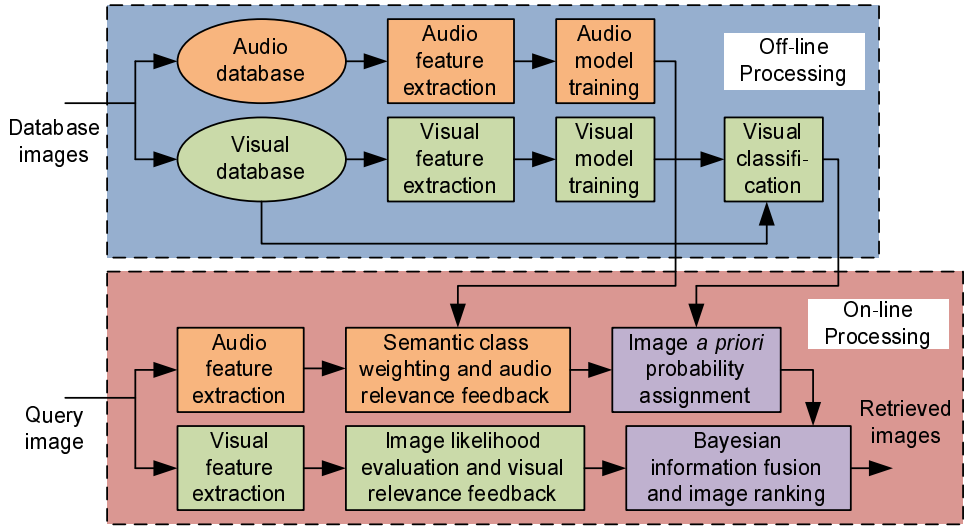


Figure 4.1: The block diagram of the proposed framework.

4.2 The Proposed Multi-modal Image Retrieval Framework

To illustrate the mechanism of the proposed framework, the diagram is divided into two parts, i.e. offline and online processing, as shown in Fig. 5.4. Both of these two parts involve visual and audio processing, which are explained with detail in what follows.

4.2.1 Processing in the Audio Domain

To effectively exploit the information in the audio domain, we borrow the popular techniques for speech recognition and speaker identification. In terms of feature selection, we employ MFCC, which has been successfully used for both speech recognition and speaker identification due to its effectiveness for characterizing the response of human auditory system. The MFCC feature of an audio clip is a sequence of vectors, where each vector is extracted from an excerpt of the clip, a.k.a. a frame. For details on the specific procedure

of the feature extraction, readers are referred to [102].

Ideally, we have an audio file for each image in the database. As an initial study, we have only an audio file for each semantic class. Suppose there are K classes, an MFCC feature sequence of length T, denoted $\boldsymbol{U}_k$=$[\boldsymbol{u}_1^k, \boldsymbol{u}_2^k, \ldots, \boldsymbol{u}_T^k]$, of the k-th class is used to train an N-state HMM, represented using $\lambda_k = \{\pi_k, A_k, B_k\}$, where $\pi_k = [\pi_1^k, \pi_2^k, \ldots, \pi_N^k]$ are the initial state probabilities, $A_k = [a_{ij}^k]_{N\times N}$ is the state transition probability matrix, $B_k = [p_1^k(\boldsymbol{u}), p_2^k(\boldsymbol{u}), \ldots, p_N^k(\boldsymbol{u})]$ are the probability density functions (PDF's) of an observation conditional on different states, and $k \in \{1, 2, \ldots, K\}$. The parameters can be estimated through the standard expectation maximization (EM) procedure. Once the HMM's are trained, the *a posteriori* probability of a semantic class given the audio features of a query can be calculated through the Bayes' theorem, i.e.

$$P(\omega_k^a|\boldsymbol{U}_q) = \frac{p(\boldsymbol{U}_q|\omega_k^a)P(\omega_k^a)}{\sum_{m=1}^{K} p(\boldsymbol{U}_q|\omega_m^a)P(\omega_m^a)}, \tag{4.1}$$

where ω_k^a denotes a class label and $p(\boldsymbol{U}_q|\omega_k^a) = p(\boldsymbol{U}_q|\lambda_k)$. These probabilities are used as the *a priori* probabilities after proper normalization, which is elaborated in the section of information fusion.

In terms of relevance feedback, we propose the following scheme. After calculating the $P(\omega_k^a|\boldsymbol{U}_q)$ for each class, a user plays and listens to the audio of the top L ones on the ranked list of the classes based on their *a posteriori* probabilities. If the relevant class appears within the examined portion of the list, the *a posteriori* probabilities given a query are recalculated for all k using the training feature of the relevant class, i.e. $P(\omega_k^a|\boldsymbol{U}_q)$=$P(\omega_k^a|\boldsymbol{U}_{TR})$, where $\boldsymbol{U}_{TR}\in\{\boldsymbol{U}_1, \boldsymbol{U}_2, \ldots, \boldsymbol{U}_K\}$ denotes the training feature of the relevant class.

4.2.2 Processing in the Visual Domain

To effectively combine the information obtained in different domains, the images are first classified based on the visual features during the offline phase. The necessity of this step can be explained as follows. Due to the fact that we only have an audio file for each semantic class, we can not directly calculate the *a posteriori* probability of each candidate image. To circumvent this problem, we choose to propagate the *a posteriori* probability of each class to the images belonging to it. To this end, the offline visual domain classification is introduced into the framework. Considering the complexity of the distribution of the visual features due to the high dimensionality, we employ a non-parametric technique for the supervised classification of images, which is known as Parzen Windows. The Bayesian decision rule is applied, which can be formulated as

$$\hat{\omega}_c^v = \underset{\omega_k^v}{\operatorname{argmax}}\, P(\omega_k^v | \boldsymbol{v}_c), \tag{4.2}$$

where ω_k^v is a class label and $\hat{\omega}_c^v$ is the class label assigned to the image, of which the visual feature is represented using $\boldsymbol{v}_c$. The $P(\omega_k^v|\boldsymbol{v}_c)$ is evaluated through the same way as in (4.1), with the audio feature replaced with the visual feature.

Before proceeding to the disucssion on the integration of the afore-mentioned information, the on-line similarity measure in the visual domain between a query and a candidate image is introduced. To this end, we employ the conventional nearest-neighbor content-based image retrieval (NN-CBIR), with the L1-norm as the distance function. In terms of relevance feedback, the query is refined based on the linear combination of the visual features of the original query, the relevant images, and the irrelevant images, a.k.a. query movement.

4.2.3 Information Fusion for Bayesian Image Audio-Visual Retrieval

As briefly introduced earlier, the overall similarity measure of the framework is based on the *a posteriori* probability of a candidate image given both the audio and visual features of a query, which can be expressed as

$$P(I_c|\boldsymbol{U}_q, \boldsymbol{v}_q) \propto p(\boldsymbol{v}_q|I_c)P(I_c|\boldsymbol{U}_q), \tag{4.3}$$

where I_c is simply the index of a candidate image.

The *a priori* probability is obtained through

$$P(I_c|\boldsymbol{U}_q) = \frac{P(\hat{\omega}_c^v|\boldsymbol{U}_q)}{\sum_i P(\hat{\omega}_i^v|\boldsymbol{U}_q)}, \tag{4.4}$$

where

$$P(\hat{\omega}_c^v|\boldsymbol{U}_q) = \begin{cases} P(\omega_k^a|\boldsymbol{U}_q), & p_{R\to q} > L \\ P(\omega_k^a|\boldsymbol{U}_{RT}), & p_{R\to q} \leq L \end{cases}, \tag{4.5}$$

where $\omega_k^a = \hat{\omega}_c^v$ and $p_{R\to q}$ the is the position of the class relevant to the query on the list ranked for audio relevance feedback. The above seemingly redundant equations can be explained as follows. The conditional probability in (4.4) can not be calculated directly in that we do not have the audio information for each image in the database. Therefore, we use the *a posteriori* probability, evaluated in the audio domain, of the semantic class to which the candidate image is classified in the visual domain, as the *a priori* probability of the candidate image in the overall framework.

In terms of the likelihood, an exponential function is used to convert the distance to

Table 4.1: Summarization of Feature Extraction.

Color Feature	
Color Histogram	8, 4, 2 bins in H, S, V channels
Color Layout	An image is partitioned into 8×8 blocks, 6, 3, 3 coefficients in Y, Cb, Cr channels
Texture Feature	
Gabor Wavelet	4 scales and 6 orientations
Shape Feature	
Fourier Descriptors	4 scales and 6 orientations

a value of a probability density function, i.e.

$$p(\boldsymbol{v}_q|I_c) = \frac{1}{A}e^{-|\boldsymbol{v}_q-\boldsymbol{v}_c|}, \tag{4.6}$$

where $A = \int e^{-|\boldsymbol{v}-\boldsymbol{v}_c|}$ is the normalizing constant.

4.3 Experiments

4.3.1 Experimental Setup

In our experimental evaluation, a collection of 4400 images featuring 44 different kinds of animals is employed as the dataset. Therefore, each animal is considered as a semantic class. The low-level feature selection is summarized in Table. 4.1. In terms of audio feature extraction, the window size (frame length) is 256 samples. In addition, the number of state N is set to 3 for the HMM's of all classes, and a Gaussian mixture with 3 components is chosen as the observation PDF given a state for each HMM. It should be noted that the parameter value selection for modeling the audio information is heuristic. In real applications, these parameters can be adjusted based on the given dataset. For experimental study, we only focus on examining the performance improvement resulting

from audio-visual information fusion. 20 images are taken from each semantic class for training the statistical model for each class in the visual domain. Another 10 images are selected from each class as the queries. We perform retrieval with no replacement since relevant images selected by the users during relevance feedback are essentially used as the training images to refine the query formulation. Regarding the evaluation criterion, we adopt two standard performance indexes, i.e. the average precision versus the number of retrieval iterations (PRI) and the precision versus recall curve (PRC). In terms of PRI, three systems are compared, including a unimodal retrieval system merely using visual features, a multimodal retrieval system employing relevance feedback only in the visual domain, and a multimodal retrieval system enabling relevance feedback in both visual and audio domain. In addition, the unimodal retrieval and multimodal retrieval with audio relevance feedback are compared based on the PRC. To facilitate the subjective evaluation, a prototype system has also been implemented.

4.3.2 Experimental Results and Analysis

Shown in Fig. 4.2 is the comparison in terms of PRI for the three systems. Users are assumed to perform relevance feedback through the inspection of top 15 classes on the ranked list in the audio domain and top 20 images on the ranked list in the visual domain. The precision is evaluated based on the top 20 images on the ranked list after each retrieval iteration. It demonstrates the improvement resulting from the integration of the information in the audio and visual domains. Meanwhile, the advantage of employing the audio relevance feedback can be observed according to the further improvement in the same figure. In Fig. 4.3, the results are shown to evaluate and compare the quality of different systems in terms of ranking the retrieved images using PRC. To

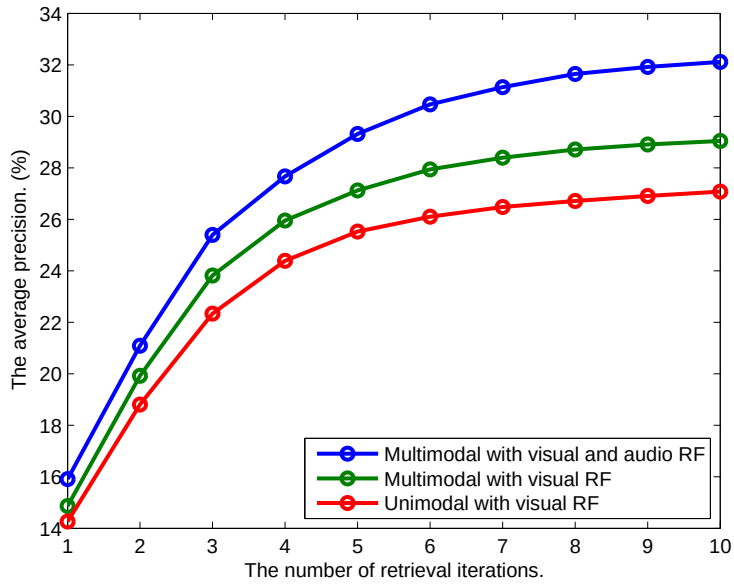


Figure 4.2: The comparison of three systems in terms of the average precision versus the number of retrieval iterations.

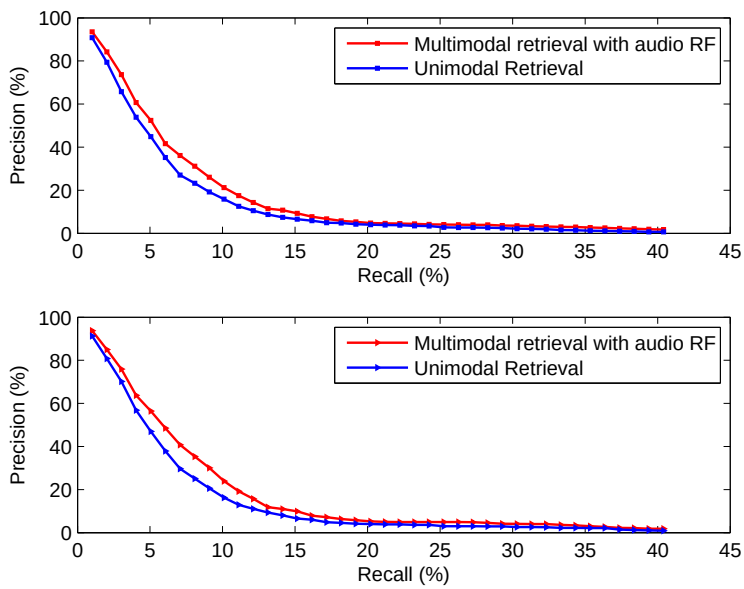


Figure 4.3: The comparison in terms of PRC between the unimodal retrieval and multimodal retrieval. Upper: result of the 5-th iteration. Lower: result of the 8-th iteration.

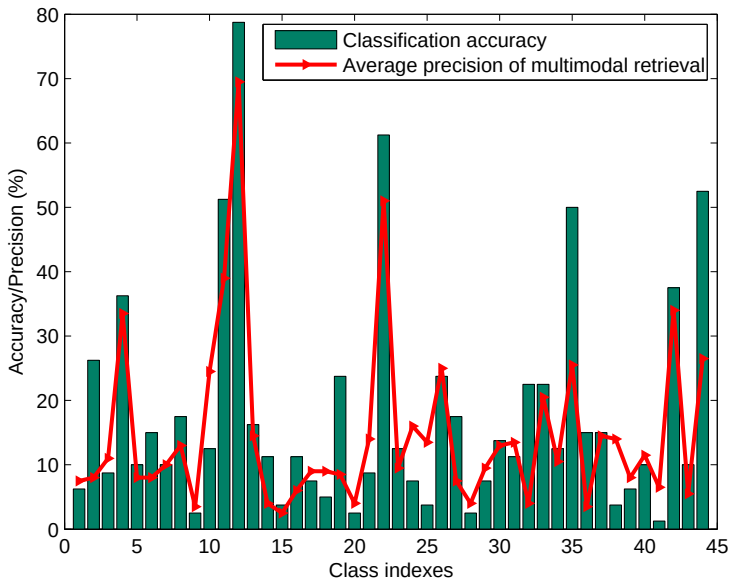


Figure 4.4: The relation between the performance of classification in visual domain and that of the retrieval.

further demonstrate the effectiveness of the proposed framework, the retrieval results of a particular query using different retrieval schemes are shown in Fig. 4.5. It can be observed that the result of the multimodal framework includes both visually similar and semantically similar images of the same class.

Apart from the above numerical evaluation, we also analyze the major performance bottleneck of the proposed system, which can be considered as a guideline for further study. As discussed earlier, the audio information can not be utilized unless it is propagated to the images in the database. The fact that the system relies on the classification in visual domain to fill this gap leads to the conjecture about the relation between the classification accuracy and the retrieval precision. To effectively demonstrate this relation, we compare the class-specific classification accuracy and the class-specific retrieval precision, and the results are shown in Fig. 4.4. The similarity between the patterns

of the fluctuation of the above two performance indexes indicates that the visual classification plays an important role in the system. The necessity of visual classification is easy to justify because currently the vast majority of the image databases do not have available audio information for each image.

4.4 Summary

Based on the Bayes' theorem, the integration of visual and audio information with the application to content-based image retrieval is studied in this work. The difference between the proposed framework and other existing ones lies in the perspective from which the information is viewed and harnessed. In addition, relevance feedback is enabled in both domains. Experimental results demonstrate the effectiveness of the developed framework in terms of both accuracy and ranking of the retrieved images. As indicated by the results, better classification approach in the visual domain is desirable in the future work.

(a) The initial retrieval result of multi-modal retrieval.

(b) The initial retrieval result of unimodal retrieval.

Figure 4.5: Subjective evaluation of the retrieval results. The filenames of the relevant images are highlighted in green.

Chapter 5

Image Annotation by Integrating Color and Texture

5.1 Introduction

To address the problem of combining low-level visual descriptors for image annotation, a new generative framework is proposed and used with the supervised classification paradigm. It combines different visual features by jointly modeling the descriptors extracted from the same salient point location of an image yet with their conditional distributions constrained via a single latent variable. In other words, to generate a set of visual words of different types of visual descriptors, the same latent visual topic is sampled and used for all of them. The input and output of the learning component are a set of image regions encompassing the particular instances of an object category and the set of parameters of the model which optimally fits the visual descriptors in the maximum likelihood sense. The input and output of the classification component are a previously unseen image region and its optimal category label in the sense of minimum probability of error. We

consider the texture and color information for inducing visual descriptors because they have proven informative and complementary for representing images capturing the semantics of general scope of real life, which are actually the targeting media content in our study. In principle, the proposed model scalable in the sense that it is capable of incorporating other kinds of visual features, such as spatial location. Details upon the selection of visual descriptors and their extraction are elaborated in a subsequent section. In terms of the specific structure of the model, it can be considered as an extension to the pLSA. Therefore, the proposed model is referred to as multi-feature pLSA (MF-pLSA) throughout the rest of the paper.

To distinguish the MF-pLSA from existing works, it can be noted first of all that such a structure avoids the vector concatenation of different visual descriptors, allowing the BOVW representations to be constructed respectively in the original descriptor spaces. Second, by assuming the statistical independence of different visual descriptors given a visual topic, their distributions are characterized using separate models yet learned jointly with the training data. Thus, it circumvents the increased dimension of the intermediate representation space, such as the case in [103]. Third, the structure of the MF-pLSA serves the purpose of modeling different types of visual descriptors more effectively than [104] in that, in our study, every descriptor of one kind has its counterpart of the other kind extracted from the region around the same key point. Moreover, compared with [34], the mixture components of the MF-pLSA is capable of dealing with intra-class variation of visual appearance. Hence, the mixture components can be referred to as visual topics as well. The contribution of the presented work is summarized as follows.

- A generative model termed MF-pLSA is proposed, which jointly models the dis-

tribution of two kinds of visual descriptor. Its advantages has been outlined as above.

- The learning algorithm of the MF-pLSA is derived based on the expectation-maximization (EM) procedure.
- As a model for supervised learning, its classification scheme is derived based on the criterion of minimum probability of classification error.
- Two databases are employed in our experimental study, i.e. VOC2009 and LabelMe. The former is a standard benchmark dataset and the latter is a dataset with a higher degree of photometric and geometric changes. The experimental study includes the comparison with histogram-based and pLSA-based approaches using vector concatenation applied at the levels of both descriptor and BOVW representation. Several performance evaluation criteria are employed, such as recall, precision, and confusion matrix.

5.2 Key Point and Visual Descriptors

Prior to the elaboration of the proposed framework for descriptor integration, we present a brief description to the visual descriptors employed in our study, which are SIFT and local transformed color histogram (LTCH). These descriptors are extracted from local regions surrounding salient points and have proven more robust against geometric change of objects, partial occlusion as well as cluttered background. The extraction of such descriptors essentially consists of two steps, which are key point detection and descriptor calculation. Given an image, a set of key points are first detected, which correspond to the locations of local image structures around which the visual information

is valuable for subsequent processing, such as object recognition, image matching and view point detection. Second, within a region around each key point, a visual descriptor is calculated based on the pixel intensity or its derivative. During this two-step process, the detection is desired to be invariant to affine transformation to local image structures and the descriptor is expected to be invariant to affine transformation of pixel colors, resulting from various kinds of illumination condition changes.

In terms of the key point detection, Mikolajczyk [56] proposed the Harris-Laplace key point detector, which is based on the Harris corner detector but improved by introducing the scale adaptiveness so as to be scale invariant. Compared with the key point detection based on difference of Gaussian proposed by Lowe [58], the Harris-Laplace detector identifies more stable key points by using the second moment matrix. It was further extended to the Harris-Affine detector, which is invariant to affine transformations of local image structures by normalizing the key point neighborhood using an estimated affine shape matrix [56]. In our present work, we only exploit the Harris-Laplace key point detector and the affine invariance is not utilized since, at the current stage, the proposed statistical modeling is by and large independent of the invariance properties of key point detection. Therefore, the evaluation and comparison among the statistical models considered in our experiments are unbiased and can be sufficiently handled by the selected detector.

The visual descriptor selection is based on the following consideration. It should be noted first that, if we consider image color or intensity as a function of pixel location, histogram-based color and SIFT descriptors are derived from the function and its first order derivative because the former and the latter characterize the information on the distributions of the intensity and the gradient of an image respectively. Hence, these two types of descriptors contain different information on the visual properties of images,

of which the proper integration may lead to some performance improvement. As for the color descriptor, since the databases used in our study include images taken under different illumination conditions, LTCH is selected due to the fact that it is invariant to light color change and shift (see [105] for the definition). This invariance property is achieved through shifting the pixel intensities in the RGB space to zero mean followed by normalizing them to unit variance. Among various histogram-based color descriptors evaluated with image and video data in [105], the LTCH produces the best results in terms of mean average precision. The SIFT descriptor is defined as the histogram of weighted gradient magnitude according to the gradient orientation. Beside the invariance property acquired by the Harris-Laplace detector, the descriptor is rotation invariant by taking the relative angle of gradient orientation with respect to the dominant gradient orientation of a key point. Moreover, it is partially invariant to light intensity change and shift because the descriptor is normalized to unit length and the gradient is based on the difference between pixel intensities. In our experiments, the Harris-Laplace detector and visual descriptor extraction implemented by Sande [105] is used (available at: http://koen.me/research/colordescriptors/).

5.3 The Proposed Framework

5.3.1 Data Generation and Representation

To view the generation of the data based on a real image, an example is shown in Fig. 5.1, where the green circles within the yellow bounding box of the boat represent the detected key points. As indicated by the highlighted circle, a descriptor can be extracted from a certain region surrounding the corresponding key point. Such descriptors from

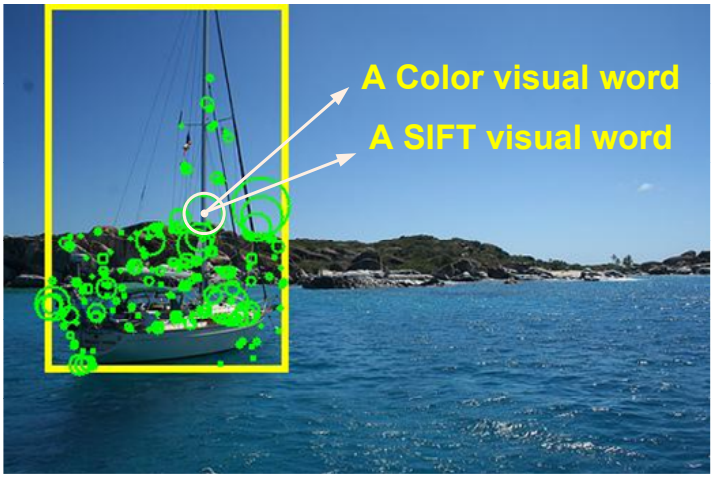


Figure 5.1: Illustration of the data obtained from an image. The yellow bounding box represents the boundary of the object, which is boat in this example. The green circles are located where there are key points detected. From each of these locations, two types of descriptors, i.e. SIFT and LTCH, are calculated.

the object regions of training images are pooled together, based on which a visual code book is built using the k-Means algorithm. As a result, the descriptor extracted at the location of each key point can be indexed using one of the code words of the visual code book, the procedure of which is referred to as visual word indexing throughout the rest of the paper. Shown in Fig. 5.1 are the output of the visual word indexing of two types of descriptors, i.e. a SIFT visual word and a LTCH visual word.

Through the above-mentioned visual word indexing procedure, we end up with a set of 3-tuples, denoted as (w_n, v_n, d_n), where $n = 1, 2, \ldots, N$, $w_n \in \{1, 2, \ldots, C\}$, $v_n \in \{1, 2, \ldots, S\}$, $d_n \in \{1, 2, \ldots, D\}$, given that there are N detected key points, C color visual words, S SIFT visual words, and D object regions. Such kind of data is defined as co-occurrence data (COD) [106] and can be summarized in a table as illustrated in Fig. 5.2, where each entry is the frequency of co-occurrence of a certain pair of SIFT and LTCH visual words and an object. By doing so, we can represent the frequency

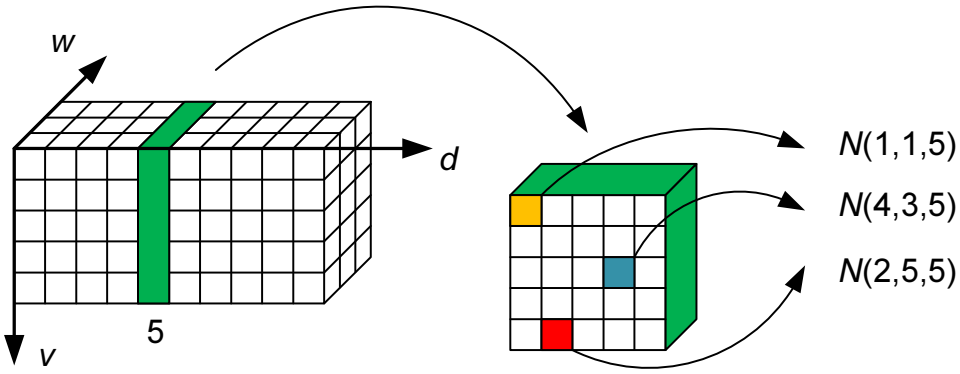


Figure 5.2: Illustration of the COD table. Slices, such as the one indicated using green color, correspond to the COD of one object region in an image. Colored blocks within the slide shown on the right-hand side represent the observed instances of COD.

of co-occurrence of the LTCH visual word w, the SIFT visual word v and the object region d as the entry of the COD table denoted by $\mathcal{N}(w, v, d)$, where $w \in \{1, 2, \ldots, C\}$, $v \in \{1, 2, \ldots, S\}$, $d \in \{1, 2, \ldots, D\}$. For example, the green slice shown in Fig. 5.2 can be thought of as the object region 4, in which only three instances of COD are observed, for $\mathcal{N}(1, 1, 5)$, $\mathcal{N}(4, 3, 5)$, $\mathcal{N}(2, 5, 5)$ times, respectively.

5.3.2 MF-pLSA for Combining SIFT and Color

The MF-pLSA is essentially a generative model, which can be illustrated using a directed graphical model shown in Fig. 5.3. The variable z is a hidden variable, of which the values are not observed along with those of (w, v, d). Compared with pLSA, there are two random variables that depend on z. In the domain of document analysis using pLSA, z is interpreted as a topic variable, which corresponds to different distributions of words when taking on different values. In the context of following discussion, we still refer to z as a topic variable but consider different values of z in the domain of visual recognition as being correspondent to a set of visual topics, which characterize the intra-class visual

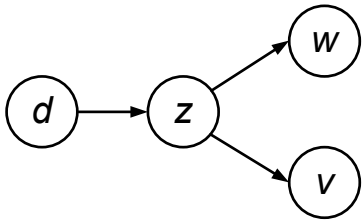


Figure 5.3: The graphical representation of the multi-feature pLSA model for combining SIFT and color descriptors.

appearance variation. From the general perspective of machine learning, it acts as a bottleneck variable which can be used to significantly reduce the number of parameters of the joint distribution of COD. Therefore, by means of linking individual components of COD with the bottleneck variable as illustrated in Fig. 5.3, over-fitting can be avoided to some extent, which is a critical feature due to the common sparsity of COD tables. Moreover, the specification of the graph structure implies the conditional independence property that the object region, the SIFT visual word and the LTCH visual word are statistically independent given z.

Based on the generative model in Fig. 5.3, the COD can also be thought of as being generated in the following process. First, to produce a tuple (w, v, d), an object region d is drawn from the distribution $P(d)$ in the first step. Second, a value of z is generated following the distribution of $P(z|d)$. Third, a SIFT visual word and a LTCH visual word are generated based on the distributions of $P(v|z)$ and $P(w|z)$, of which the order is not considered in the procedure. According to this data generation process, the joint probability mass function $P(w, v, d)$ modeled using MF-pLSA can be formulated in the

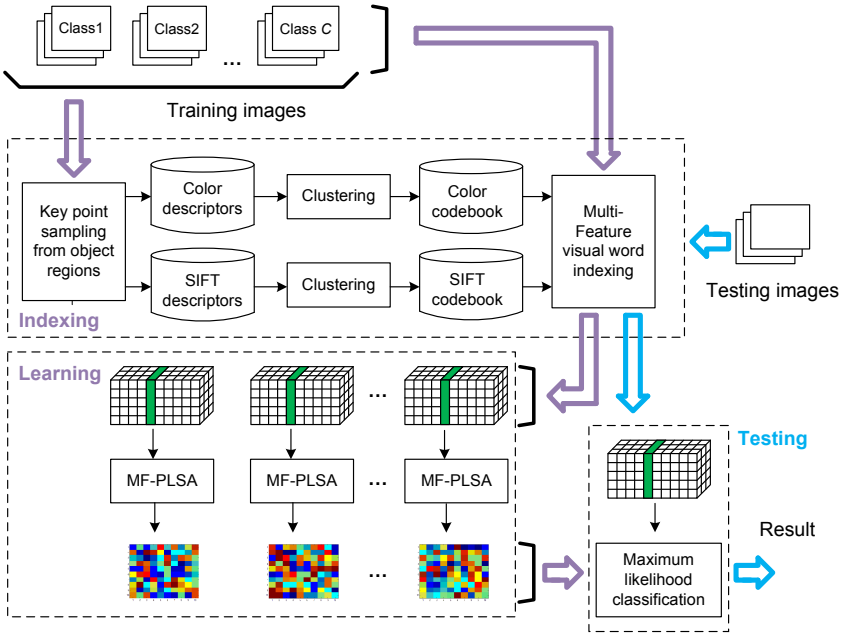


Figure 5.4: The block diagram of the proposed framework for image annotation by integrating color and SIFT descriptors.

parametric form of

$$\begin{aligned} P(w,v,d) &= P(d)P(w,v|d) \\ &= P(d)\sum_z P(w|z)P(v|z)P(z|d), \end{aligned} \tag{5.1}$$

which is a mixture of the distributions of visual words conditional on the topic variable. The MF-pLSA is used in a framework of supervised classification, where a joint distribution of SIFT and LTCH visual words is learned for each class using the training samples belonging to that class. To keep the notation uncluttered, the elaboration of the learning algorithms of the MF-pLSA is made independent of the class identity because the training algorithm remains identical for all classes.

5.3.3 The Learning Algorithm of MF-pLSA

In view of clarity, we denote the set of parameters of the model as Θ, where $\Theta = \Theta_{w,v|d} \cup \Theta_d$ and

$$\begin{aligned} \Theta_{w,v|d} &\triangleq \{P(w|z)|(w,z) \in \{1,2,\ldots,S\} \times \{1,2,\ldots,Z\}\} \\ &\cup \{P(v|z)|(v,z) \in \{1,2,\ldots,S\} \times \{1,2,\ldots,Z\}\} \\ &\cup \{P(z|d)|(z,d) \in \{1,2,\ldots,Z\} \times \{1,2,\ldots,D\}\}, \end{aligned} \tag{5.2}$$

where Z is the number of mixture components, and $\Theta_d \triangleq \{P(d)|d \in \{1,2,\ldots,D\}\}$. Defining the training data as $\mathbf{T} = \{(w_n, v_n, d_n)|n = 1,2,\ldots,N\}$, the goal of learning the MF-pLSA is to find the set of parameters Θ such that the model can account for $\mathbf{T}$ most appropriately based on a chosen criterion. To this end, we adopt the maximum likelihood principle for learning the MF-pLSA model. Based on (5.1), the log likelihood of $\mathbf{T}$ with respect to the model Θ can be written as

$$\begin{aligned} L(\mathbf{T}|\Theta) &= \sum_{d=1}^{D}\sum_{w=1}^{W}\sum_{v=1}^{V} \mathcal{N}(w,v,d) \log P(w,v,d) \\ &= N\sum_{d=1}^{D} \tilde{P}(d) \log P(d) + \sum_{d=1}^{D}\sum_{w=1}^{W}\sum_{v=1}^{V} \mathcal{N}(w,v,d) \log \sum_{z=1}^{Z} P(w|z)P(v|z)P(z|d). \end{aligned} \tag{5.3}$$

Since $\Theta_{w,v|d}$ and Θ_d are separately involved with the two terms in (5.3) and $\Theta_{w,v|d} \bigcap \Theta_d = \phi$, the two terms can be maximized independently.

Based on the definitions of entropy $H(\cdot)$ and relative entropy $D(\cdot||\cdot)$, we have

$$\begin{aligned} D(\tilde{P}(d)||P(d)) &= \sum_{d=1}^{D} \tilde{P}(d) \log \frac{\tilde{P}(d)}{P(d)} \\ &= -\sum_{d=1}^{D} \tilde{P}(d) \log P(d) - H\left(\tilde{P}(d)\right). \end{aligned} \tag{5.4}$$

It can be observed that the maximization of the first term in (5.3) amounts to the minimization of $D(\tilde{P}(d)||P(d))$. Since $D(\tilde{P}(d)||P(d)) = 0$ if and only if $\tilde{P}(d) = P(d)$, the distribution of the object region is its empirical distribution.

In terms of the second term in (5.3), as the hidden variable z is not observed, the model is learned using incomplete data. Therefore, the expectation-maximization (EM) procedure is one ideal tool to estimate the parameters $\Theta_{w,v|d}$ associated with the second term. Let $\boldsymbol{z} = \{z_n | n = 1, 2, \ldots, N\}$ be a set of random variables and assume they are observed, the complete log likelihood of $\mathbf{T} \bigcup \boldsymbol{z}$ with respect to the model $\Theta_{w,v|d}$, denoted $L(\mathbf{T}, \boldsymbol{z}|\Theta_{w,v|d})$, can be formulated as

$$\begin{aligned} L(\mathbf{T}, \boldsymbol{z}|\Theta_{w,v|d}) &= \sum_{n=1}^{N} \log P(w_n, v_n, z_n, d_n) \\ &= \sum_{n=1}^{N} \log P(w_n|z_n) P(v_n|z_n) P(z_n|d_n), \end{aligned} \tag{5.5}$$

which is a function of $\boldsymbol{z}$ and hence a random variable as well. The expectation of this random variable is calculated in the E-step followed by the maximization of this expectation in the M-step.

E-Step

As usual, the objective of the E-Step is to derive the expectation of $L(\mathbf{T}, \boldsymbol{z}|\Theta_{w,v|d})$ with respect to the *a posteriori* probability $P(\boldsymbol{z}|\mathbf{T}, \Theta_{w,v|d}^{<i-1>})$[1], where $\Theta_{w,v|d}^{<i-1>}$ is the set of parameters resulting from the $(i-1)$-th iteration of the EM procedure. According to the assumption on the structure of the model, we have

$$P(\boldsymbol{z}|\mathbf{T}, \Theta_{w,v|d}^{<i-1>}) = \prod_{n=1}^{N} \frac{P^{<i-1>}(w_n|z_n)P^{<i-1>}(v_n|z_n)P^{<i-1>}(z_n|d_n)}{\sum_{z_n=1}^{Z} P^{<i-1>}(w_n|z_n)P^{<i-1>}(v_n|z_n)P^{<i-1>}(z_n|d_n)}. \tag{5.6}$$

The expectation, a.k.a. the Q-function, can be calculated by

$$\begin{aligned} Q(\Theta_{w,v|d}^{<i>}, \Theta_{w,v|d}^{<i-1>}) &= \sum_{\boldsymbol{z}} \sum_{n=1}^{N} \log P(w_n|z_n)P(v_n|z_n)P(z_n|d_n)P(\boldsymbol{z}|\mathbf{T}, \Theta_{w,v|d}^{<i-1>}) \\ &= \sum_{z=1}^{Z} \sum_{d=1}^{D} \sum_{w=1}^{W} \sum_{v=1}^{V} \mathcal{N}(w,v,d)P(z|w,v,d,\Theta_{w,v|d}^{<i-1>}) \log P(w|z)P(v|z)P(z|d). \end{aligned} \tag{5.7}$$

M-Step

The optimization in the M-Step can be formulated as

$$\begin{aligned} \text{maximize} \quad & Q(\Theta_{w,v|d}^{<i>}, \Theta_{w,v|d}^{<i-1>}) \\ \text{subject to:} \quad & \sum_{w=1}^{W} P(w|z) = 1, z = 1, 2, \ldots, Z \\ & \sum_{v=1}^{V} P(v|z) = 1, z = 1, 2, \ldots, Z \\ & \sum_{z=1}^{Z} P(z|d) = 1, d = 1, 2, \ldots, D \end{aligned} \tag{5.8}$$

[1]We use the brackets to distinguish the index of iteration from the exponent.

We maximize the $Q(\Theta^{<i>}_{w,v|d}, \Theta^{<i-1>}_{w,v|d})$ with respect to $\Theta^{<i>}_{w,v|d}$ by introducing Lagrangian multipliers μ_z, ν_z and λ_d. Accordingly, the resulting lagrangian function has the form

$$\begin{aligned}
&\Lambda(\Theta^{<i>}_{w,v|d}, \boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\lambda}) \qquad (5.9)\\
&= \sum_{z=1}^{Z}\sum_{d=1}^{D}\sum_{w=1}^{W}\sum_{v=1}^{V}\mathcal{N}(w,v,d)P(z|w,v,d,\Theta^{<i-1>}_{w,v|d})\log P(w|z)P(v|z)P(z|d)\\
&+ \sum_{z=1}^{Z}\mu_z\left(\sum_{w=1}^{W}P(w|z)-1\right)+\sum_{z=1}^{Z}\nu_z\left(\sum_{v=1}^{V}P(v|z)-1\right)\\
&+ \sum_{d=1}^{D}\lambda_d\left(\sum_{z=1}^{Z}P(z|d)-1\right),
\end{aligned}$$

where $\boldsymbol{\mu} = \{\mu_z | z = 1, 2, \ldots, Z\}$, $\boldsymbol{\nu} = \{\nu_z | z = 1, 2, \ldots, Z\}$ and $\boldsymbol{\lambda} = \{\lambda_d | d = 1, 2, \ldots, D\}$. Solving the follow equation system for $\Theta^{<i>}_{w,v|d}$,

$$\begin{cases}
\partial Q(\Theta^{<i>}_{w,v|d}, \Theta^{<i-1>}_{w,v|d})/\partial P(w|z) = 0\\
\partial Q(\Theta^{<i>}_{w,v|d}, \Theta^{<i-1>}_{w,v|d})/\partial P(v|z) = 0\\
\partial Q(\Theta^{<i>}_{w,v|d}, \Theta^{<i-1>}_{w,v|d})/\partial P(z|d) = 0
\end{cases},$$

we end up with the following estimates of the parameters,

$$\begin{aligned}
P(w|z) &= \frac{\sum_{d=1}^{D}\sum_{v=1}^{V}\mathcal{N}(w,v,d)P(z|w,v,d,\Theta^{<i-1>}_{w,v|d})}{\sum_{d=1}^{D}\sum_{w=1}^{W}\sum_{v=1}^{V}\mathcal{N}(w,v,d)P(z|w,v,d,\Theta^{<i-1>}_{w,v|d})}, \qquad (5.10)\\
P(v|z) &= \frac{\sum_{d=1}^{D}\sum_{w=1}^{W}\mathcal{N}(w,v,d)P(z|w,v,d,\Theta^{<i-1>}_{w,v|d})}{\sum_{d=1}^{D}\sum_{w=1}^{W}\sum_{v=1}^{V}\mathcal{N}(w,v,d)P(z|w,v,d,\Theta^{<i-1>}_{w,v|d})},\\
P(z|d) &= \frac{\sum_{w=1}^{W}\sum_{v=1}^{V}\mathcal{N}(w,v,d)P(z|w,v,d,\Theta^{<i-1>}_{w,v|d})}{\mathcal{N}(d)},
\end{aligned}$$

where $\mathcal{N}(d) = \sum_{w=1}^{W}\sum_{v=1}^{V}\mathcal{N}(w,v,d)$.

5.3.4 Supervised Classification with MF-pLSA

To annotate object regions of images, we apply supervised classification using MF-pLSA. Define the set of class labels as $\Omega = \{\omega|1, 2, \ldots, |\Omega|\}$, where $|\Omega|$ is the number of classes. The joint distribution of SIFT and LTCH visual words for class ω can be calculated by marginalizing out the region variable d through $P_\omega(w, v) = \sum_{d=1}^{D_\omega} P_\omega(w, v, d)$, where D_ω is the number of object regions used to estimate the $P_\omega(w, v, d)$, the joint distribution of the visual words and object regions of class ω modeled using MF-pLSA. The maximum likelihood principle is employed to classify a previously unseen object region, denoted d'. Let $\mathbf{T}'$ be the set of COD of d', the log likelihood of $\mathbf{T}'$ with respect to the model of class ω, denoted Θ_ω, can be calculated via

$$\begin{aligned} L(\mathbf{T}'|\Theta_\omega) &= \sum_{w=1}^{W}\sum_{v=1}^{V} \mathcal{N}(w, v, d') \log P_\omega(w, v) \\ &= \sum_{w=1}^{W}\sum_{v=1}^{V} \mathcal{N}(w, v, d') \log \sum_{d=1}^{D_\omega} P_\omega(d) \sum_{z=1}^{Z} P_\omega(w|z) P_\omega(v|z) P_\omega(z|d), \end{aligned} \tag{5.11}$$

where $P_\omega(d)$, $P_\omega(w|z)$, $P_\omega(v|z)$ and $P_\omega(z|d)$ are the parameters of the MF-pLSA of the ω-th class. The decision on the class label is made by

$$\hat{\omega}' = \arg\max_{\omega \in \Omega} L(\mathbf{T}'|\Theta_\omega). \tag{5.12}$$

It should be noted that there is generally no *a priori* knowledge on the distribution of object categories, which means the uniform distribution is a reasonable assumption. In this case, the maximum likelihood classification is equivalent to the maximum *a posteriori* probability classification. Therefore, the result of the above classification scheme using the MF-pLSA model is optimal in the sense of minimum probability of error.

5.4 Experiments

5.4.1 Databases

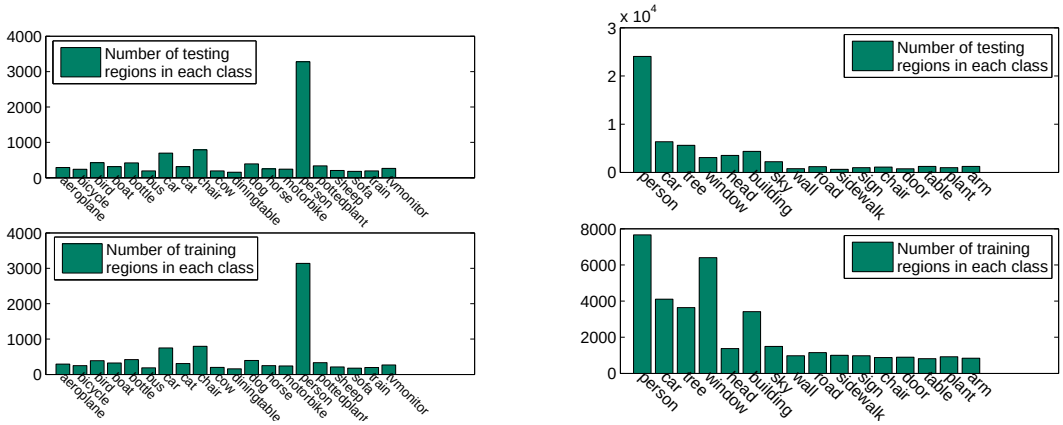
We employ two databases for studying the performance of the multi-feature pLSA model, i.e. the VOC2009 [107] and the LabelMe [108] databases. The two databases are summarized as follows and the sizes of training and testing sets of the above two databases are summarized in Fig. 5.5.

The VOC2009 Database

The database includes 7818 images in total, which are annotated using 20 concepts. All of the 20 classes are selected for the experiments. The resulting training and testing sets include 9298 and 9390 object regions, respectively. In this database, the boundaries of objects are only manually outlined with rectangles.

The LabelMe Database

The content of the LabelMe database depends on when it is downloaded because it is continuously updated as interested users annotate the images. The version of LabelMe database we downloaded includes 51285 images and 306005 regions in total, which are annotated using 210 concepts, where each concept is represented as a keyword and considered as a class. Only 16 classes are selected for the experiments because they have significantly more annotated object regions than those unselected do. The images of each of the 16 classes are randomly split into two disjoint sets to prepare the training and testing data. As a result, the training set and the testing set include 36547 and 58176 regions, respectively. Moreover, the boundaries of objects have been roughly outlined manually with polygons in the images of this database.



(a) VOC2009 database

(b) LabelMe database

Figure 5.5: The numbers of training and testing samples of different classes within the two employed databases.

5.4.2 Experimental Setup

Approaches Compared with the MF-pLSA

To evaluate the performance of the multi-feature pLSA model, we compare it with seven types of approaches to the same classification task, including the histogram of SIFT visual word (S-Hist), the histogram of LTCH visual word (C-Hist), the histogram of visual word of combined SIFT and LTCH descriptor (SC-Hist), the concatenated S-Hist and C-Hist (Concat-Hist), the pLSA of SIFT visual word (S-pLSA), the pLSA of LTCH visual word (C-pLSA) and the pLSA of combined SIFT and color descriptor (SC-pLSA). In fact, the SC-Hist, Concat-Hist and SC-pLSA integrate the information from the SIFT and LTCH descriptors by different means. To be specific, the SC-Hist can be considered as a low-level fusion of the two kinds of descriptors as the two individual descriptor vectors are directly concatenated to form a longer vector, i.e. the descriptor vector concatenation. On the other hand, the Concat-Hist combines the S-Hist and C-Hist into a longer histogram, which can be thought of as the fusion at an intermediate level of image representation,

i.e. the BOVW vector concatenation. Similarly, the SC-pLSA is based upon the same sort of descriptor fusion as that of SC-Hist but with a different modeling strategy.

Implementation

The entire framework, from preparing for the datasets, through preprocessing, such as learning visual codebooks for different descriptors, to learning and classification, is implemented using C++ with the objected-oriented programming paradigm. The implementation includes a basic I/O static library mainly for accessing the data and results involved in the experiments, a dynamically linked library for building codebooks considering its extensive usage in most of the BOVW-based image understanding tasks, and a static library including the learning and classification algorithms of the proposed model and all others that are compared. The softwares were built for both MS Windows and GNU Linux platforms and data parallelization is enabled to speed up the calculation. The design of this implementation is for the purpose of efficient computing and sustainable development of a statistical pattern recognition software toolkit. Training samples, testing samples as well as the classification results are accessible throught the web site at http://clbir.rml.ryerson.ca/maindisplay.html.

Parameters of the model, such as the size of a codebook and the number of visual topics of a pLSA-based model, are set as follows. For the approaches based on mixture modeling, such pLSA and multi-feature pLSA, we set the number of hidden topics Z to 5, 20 and 50 in order to study the performance variation as the model order changes, which addresses the characterization of intra-class variation of visual appearance. To test how the level of finess of descriptor quantization affects the overall performance, we conduct experiments with 100, 200 and 500 visual words. In all cases of the comparative study, the numbers of visual words for different descriptors are kept equal.

Performance Evaluation Criteria

In terms of performance evaluation criterion, we first adopt class-specific recall, denoted $\mathcal{R}_\omega$, and precision, denoted $\mathcal{P}_\omega$, defined respectively as

$$\mathcal{R}_\omega = \frac{R_\omega^{corr}}{R_\omega} \tag{5.13}$$

and

$$\mathcal{P}_\omega = \frac{R_\omega^{corr}}{R_\omega^{targ}}, \tag{5.14}$$

where R_ω^{corr}, R_ω and R_ω^{targ} denote the number of correctly classified regions of the ω-th class, the number of regions of the ω-th class, and the number of regions categorized into the ω-th class. Second, we consider the average recall and precision, defined by

$$\mathcal{R}_\omega^{avg} = \frac{1}{|\Omega|} \sum_\omega \mathcal{R}_\omega \tag{5.15}$$

and

$$\mathcal{P}_\omega^{avg} = \frac{1}{|\Omega|} \sum_\omega \mathcal{P}_\omega. \tag{5.16}$$

Furthermore, to examine the details of misclassification, we employ confusion matrix. Finally, to visually inspect the results of annotation, some examples of the classification using SC-pLSA and MF-pLSA are illustrated in Fig. 5.22. It should be noted that the purpose of the first two is to compare the proposed approach with others whereas the last two are intended to find out which classes are difficult to be distinguished from each other based on numerical results and inspecting the visual appearance of the samples.

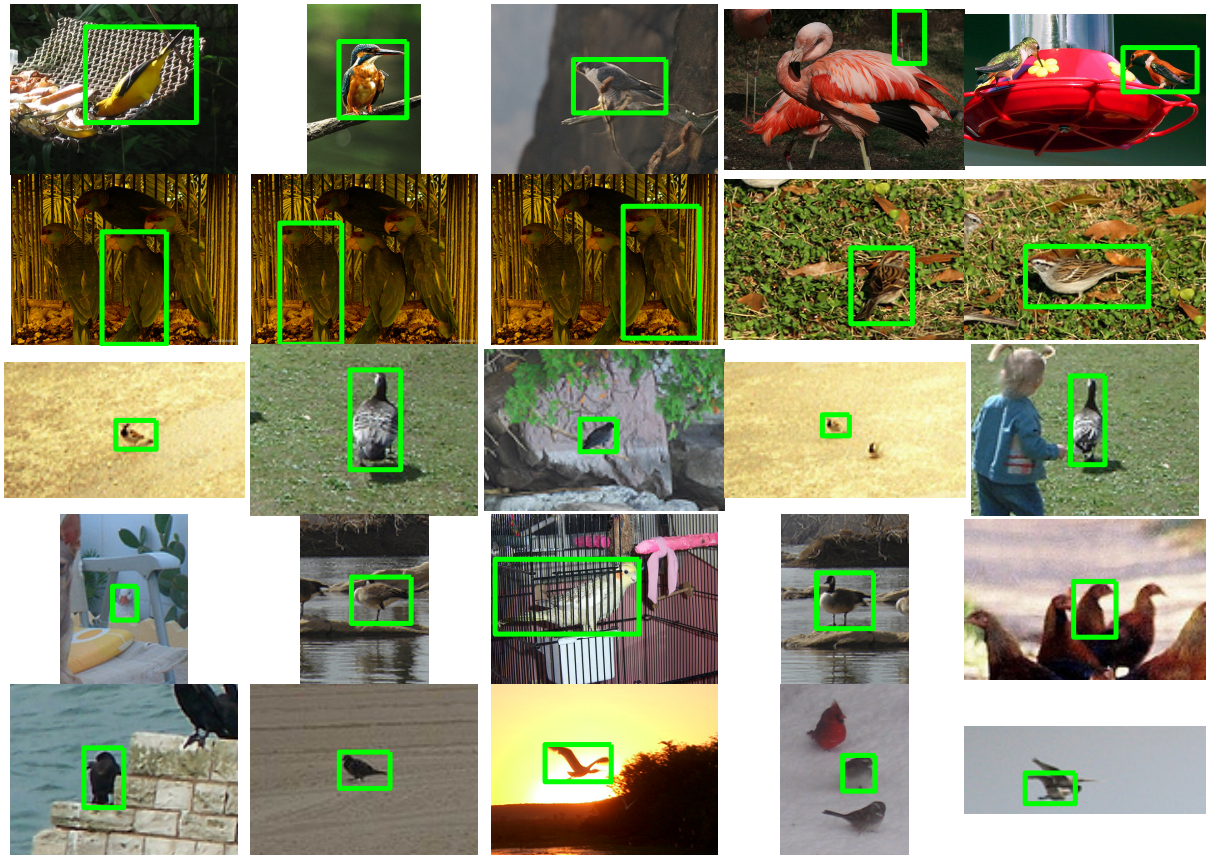


Figure 5.6: 5 visual topics from the bird class of the VOC2009 database. Each row includes the first 5 region of a certain the visual topic ranked based on the *a posteriori* probability of the visual topic given the region , i.e. $P(z|d)$. For better illustration, the images are independently scaled such that the regions of interest fit the area of display while maintaining the quality for visually recognizing the objects.

5.4.3 Experimental Results

We first illustrate two examples of the learned visual topics. Shown in Fig. 5.6 and Fig. 5.7 are the 5 visual topics learned using MF-pLSA for the class of tree of the LabelMe database and the class of bird of the VOC2009 databases, respectively. These topics are learned with $Z = 5$. For each of the visual topics, the regions are ranked based on the $P(z|d)$ in descending order. The first 5 regions on the ranked list of a topic are selected for visualizing the appearance of the topic. The idea is that for a given topic

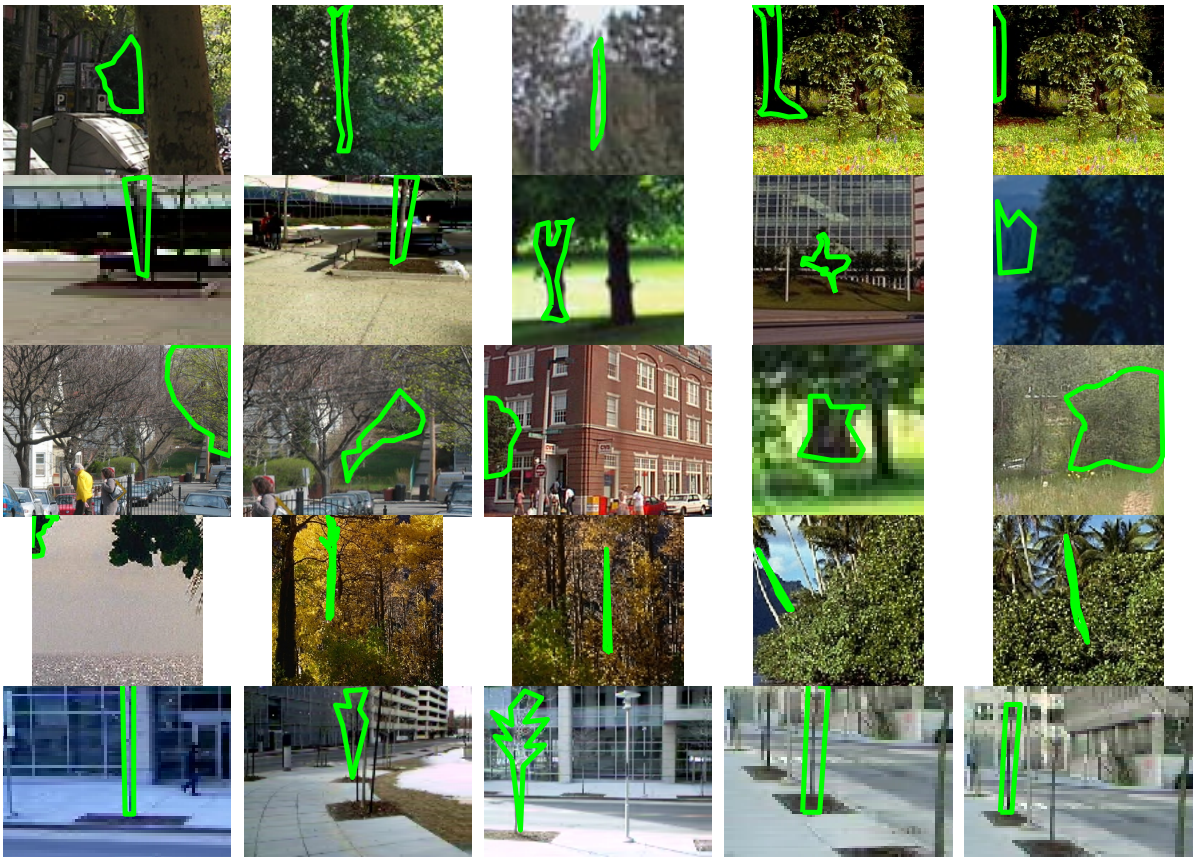


Figure 5.7: 5 visual topics from the tree class of the LabelMe database. Each row includes the first 5 region of a certain the visual topic ranked based on the *a posteriori* probability of the visual topic given the region , i.e. $P(z|d)$. For better illustration, the images are independently scaled such that the regions of interest fit the area of display while maintaining the quality for visually recognizing the objects.

z, the higher the $P(z|d)$ of a region is compared with those of the other regions, the more the region is generated by selecting the topic z. This in turn leads to the fact that the visual appearance of topic z is more easily observed from such a region as d. Since each topic corresponds to the distribution of both SIFT and LTCH visual words, different combinations of color and texture visual patterns are expected to be observed. In Fig. 5.6, the first three visual topics contain rich texture but their colors are different. By contrast, the other two visual topics do not have much texture while their colors

are also different. In Fig. 5.7, the third visual topic corresponds to the visual pattern of small branches and leaves of a tree, whereas the others primarily characterize trunks with different color and shape patterns.

In terms of the study on classification performance, the results shown in Table 5.1 and Table 5.2 are the class-specific recall and precision evaluated using the VOC2009 database. The number of visual word is 500. The maximum recall and precision values of each class are highlighted. In terms of recall, we obtain improved performance for 12 out of 20 classes, whereas, in terms of precision, better performance is observed for 15 of the 20 classes. Shown in Table 5.3 and Table 5.4 are the numerical results of class-specific recall and precision evaluated using the LabelMe database with 200 visual words. Out of 16 classes, the recall values of 11 classes and the precision values of 10 classes get improved using the MF-pLSA model. To efficiently utilize the available space for presenting the results obtained with different number of visual words and visual topics, average recall and precision are employed, as shown in Fig. 5.8 and Fig. 5.9. For the VOC2009 database, according to Fig. 5.8(a) through Fig. 5.8(f), about 3% improvement in terms of $\mathcal{R}_{\omega}^{avg}$ is obtained compared with the best of the other approaches. In terms of $\mathcal{P}_{\omega}^{avg}$, the improvement falls between 1.5% and 2.0%, increasing as the number of visual words is raised. As shown in Fig. 5.9(a) through Fig. 5.9(f), improvement in terms of $\mathcal{R}_{\omega}^{avg}$ and $\mathcal{P}_{\omega}^{avg}$ for the LabelMe database range from 4.4% to 6.5% and 3.5% to 5.2%, respectively. It can be observed that the performance based on the LabelMe database is better than that based on the VOC2009 database under different experimental settings.

It is also observed from Fig. 5.8 and Fig. 5.9 that, for different numbers of hidden topics, we end up with very similar results using the pLSA-based approaches, i.e. S-pLSA, C-pLSA and SC-pLSA. This phenomenon directly results from using the pLSA

Table 5.1: The class-specific recall of the VOC2009 database using 500 visual words.

	S-Hist	C-Hist	SC-Hist	Concat-Hist	S-pLSA	C-pLSA	SC-pLSA	MF-pLSA 5	MF-pLSA 20	MF-pLSA 50
Aeroplane	61.94	47.4	5.88	58.82	67.82	61.59	68.86	**71.63**	69.55	64.01
Bicycle	24.17	12.08	2.08	14.17	**41.25**	23.33	40.42	37.5	40.00	37.5
Bird	**12.62**	2.1	1.17	3.5	8.88	3.97	8.88	9.81	8.41	10.05
Boat	20.25	18.04	6.65	18.35	15.51	26.58	12.03	**32.59**	26.27	23.73
Bottle	29.83	12.17	10.98	23.63	**39.86**	14.08	37.95	18.14	11.69	8.35
Bus	60.31	14.95	3.61	38.66	**61.86**	37.63	61.34	58.76	55.67	59.79
Car	39.94	16.67	1.15	35.34	34.63	14.8	37.21	40.95	**43.97**	43.1
Cat	44.44	**77.46**	4.13	3.81	36.19	28.25	36.83	36.19	39.37	43.17
Chair	9.49	1.52	1.9	3.42	14.43	4.18	15.57	19.24	19.62	**20.38**
Cow	4.66	4.15	5.7	4.66	6.22	9.33	6.22	**12.95**	5.7	5.7
Dining table	2.56	2.56	3.85	4.49	17.95	22.44	16.03	**26.92**	19.87	19.23
Dog	6.62	4.07	0.25	2.04	23.92	9.16	23.92	**30.53**	26.21	30.28
Horse	21.03	3.97	3.17	3.97	25.79	14.29	25.00	**30.16**	28.17	28.97
motorbike	9.05	32.51	2.06	32.1	39.09	23.05	37.86	42.39	**43.21**	38.27
Person	5.43	0.58	0.27	1.89	26.99	9.48	24.95	34.68	45.2	**52.42**
Potted plant	7.42	27.89	3.26	**32.64**	13.06	29.08	12.17	27.00	21.66	20.18
Sheep	14.01	20.77	10.63	**72.95**	21.74	20.29	26.57	14.01	11.59	9.66
Sofa	3.83	3.28	3.83	5.46	14.21	14.21	15.3	15.3	**16.94**	15.3
Train	2.06	2.06	2.06	4.64	34.54	16.49	33.51	37.63	**43.81**	43.3
TV monitor	**61.65**	19.17	6.77	40.98	42.86	27.44	41.73	36.09	25.56	18.42

Table 5.2: The class-specific precision of the VOC2009 database using 500 visual words.

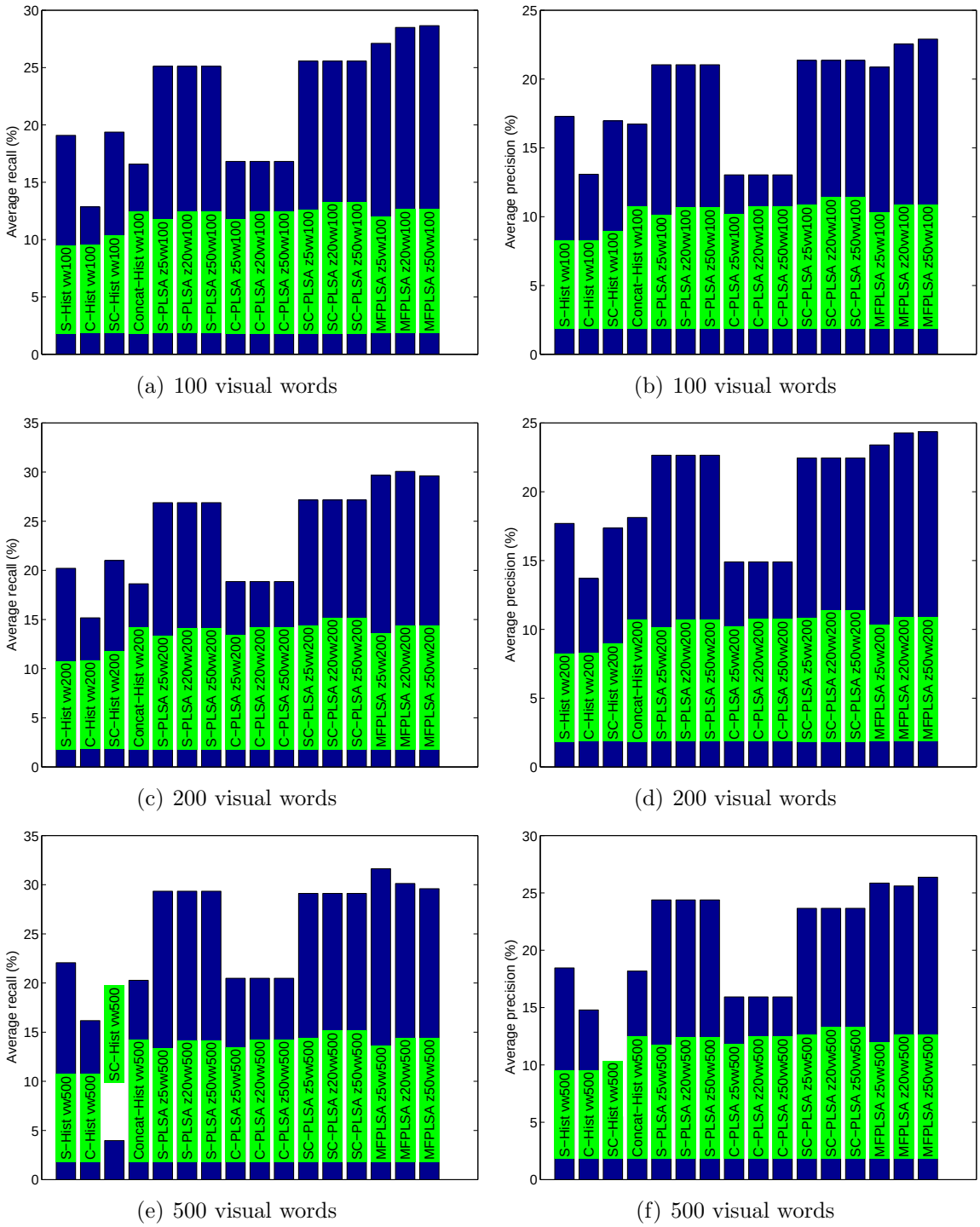
	S-Hist	C-Hist	SC-Hist	Concat-Hist	S-pLSA	C-pLSA	SC-pLSA	MF-pLSA 5	MF-pLSA 20	MF-pLSA 50
Aeroplane	18.04	23.66	1.22	30.36	23.39	21.02	25.61	29.57	**30.55**	31.57
Bicycle	26.13	8.5	3.36	13.71	**43.42**	9.66	38.96	26.71	27.51	28.04
Bird	13.14	9.78	2.23	17.86	20.99	16.83	22.49	**31.34**	25.9	25.15
Boat	15.27	16.33	2.43	17.21	27.68	14.76	19.9	**25.81**	25.3	23.81
Bottle	9.58	17.29	4.33	15.11	13.46	13.23	13.89	20.88	23.22	**26.12**
Bus	18.28	12.13	1.81	15.92	22.64	12.05	22.5	25	24.83	**25.05**
Car	34.79	20.9	3.29	42.2	41.27	21.02	40.72	**42.99**	42.15	42.25
Cat	21.05	7.97	3.28	12.24	24.41	20.18	26.07	**31.84**	30.02	31.26
Chair	24.51	**35.29**	4.27	27.84	23.9	13.64	23.7	25.76	30.69	33.75
Cow	6.29	5.16	2.55	2.74	7.1	8.91	7.36	**14.79**	11.83	13.92
Dining table	5.8	3.51	1.84	**8.33**	7.29	4.62	6.41	8.24	6.7	7.41
Dog	19.55	5.97	1.28	5.8	23.92	14.81	23.98	19.74	20.6	**24.09**
Horse	12.47	4.46	4.23	5.38	20	12.59	19.75	17.97	18.93	**20.22**
Motorbike	11.83	7.72	3.07	14.44	14.77	6.97	14.51	**15.01**	14.64	13.74
Person	**76.72**	70.37	60	71.26	71.31	70.36	71.07	74.07	68.26	65.46
Potted plant	14.97	15.21	2.39	15.19	12.05	18.7	9.95	22.81	21.22	**23.29**
Sheep	2.99	4.91	1.48	4.34	11.72	13.86	11.78	15.85	16.11	**19.61**
Sofa	11.67	8.00	2.87	8.93	11.35	5.42	11.38	9.79	12.25	**12.28**
Train	12.5	5.97	1.53	13.43	31.6	9.38	**30.95**	22.26	21.85	21.71
TV monitor	13.5	12.78	2.72	21.46	35.4	10.64	32.08	36.64	**39.53**	38.58

Table 5.3: The class-specific recall of the LabelMe database using 200 visual words.

	S-Hist	C-Hist	SC-Hist	Concat-Hist	S-pLSA	C-pLSA	SC-pLSA	MF-pLSA 5	MF-pLSA 20	MF-pLSA 50
Person	31.3	4.46	27.58	20.55	70.23	30.4	69.48	70.29	75.49	**76.79**
Car	53.33	27.24	56.63	49.52	47.8	43.2	48.94	58.95	**60.67**	60.64
Tree	27.08	47.68	31.7	**53.39**	29.89	28.59	29.19	32.59	33.7	34.6
Window	39.18	18.44	40.7	38.89	42.97	16.1	41.87	36.88	**43.78**	43.03
Head	**32.76**	8.96	31.99	24.99	26.86	15.37	27.03	28.05	25.3	21.1
Building	0.32	0.05	0.11	0.46	12.13	8.6	12.89	22.38	25.02	**28.98**
Sky	54.63	62.72	62.54	65.29	38.98	57.96	49.6	63.49	66.64	**68.12**
Wall	2.66	1.27	2.54	6.6	16.37	8.38	16.62	23.22	27.28	**27.79**
Road	**61.09**	42.26	58.33	58.33	52.72	37.66	50.79	52.8	52.8	54.56
Sidewalk	33.49	35.34	37.96	40.43	33.18	42.28	37.65	47.53	**50.77**	47.99
Sign	17.65	5.98	16.94	13.59	**28.6**	22.92	23.53	28.5	26.98	24.85
Chair	2.74	13.79	2.56	11.05	12.33	13.42	10.96	**15.98**	14.7	12.51
Door	38.08	1.84	37.15	29.78	30.43	5.14	33.33	**38.6**	30.3	25.82
Table	2.82	7.91	1.86	9.52	18.48	18	17.68	36.08	36.08	**37.05**
Plant	8.04	37.47	11.51	34.62	29.23	40.33	34.93	43.48	45.42	**46.84**
Arm	**46.96**	7.53	46.07	34.09	31.9	6.96	28.99	16.28	12.23	11.26

Table 5.4: The class-specific precision of the LabelMe database using 200 visual words.

	S-Hist	C-Hist	SC-Hist	Concat-Hist	S-pLSA	C-pLSA	SC-pLSA	MF-pLSA 5	MF-pLSA 20	MF-pLSA 50
Person	93.96	78.02	**95.14**	95.2	91.69	87.15	92.13	91.64	91.07	90.78
Car	57.8	28.76	56.08	62.45	57.56	27.12	56.65	61.54	62.34	**62.48**
Tree	52.09	23.96	50.04	29.71	41.81	43.06	39.98	52.61	**52.98**	52.9
Window	26.35	31.49	25.13	36.34	40.18	17.59	38.62	45.12	47.39	**47.5**
Head	24.48	11.21	27.82	20.97	31.46	20.65	30.87	36.77	40.75	**42.64**
Building	**82.35**	3.13	50	54.05	42.32	18.35	38.21	31.39	31.7	32.79
Sky	21.55	28.47	27.75	33.03	28.82	32.06	38.78	42.98	44.04	**44.56**
Wall	6.56	8	6.41	**10.22**	8	6.33	7.82	7.82	8.87	8.88
Road	21.76	11.35	25.03	22.54	19.35	16.1	20.95	29.12	31.58	**31.64**
Sidewalk	7.92	3.8	8.54	7.79	15.25	7.18	14.06	16.39	**17.04**	16.67
Sign	11.13	10.5	9.96	18.03	11.04	5.94	11.61	16.46	19.19	**20.45**
Chair	12.66	5.84	9.62	**18.42**	11.07	3.67	10.7	9.82	12.8	12.4
Door	9.24	8.54	8.25	**14.05**	8.92	2.34	9.19	11.52	14.02	13.26
Table	13.16	13.07	13.37	**34.1**	15.05	12.81	15.2	23.37	24.51	23.87
Plant	11.52	2.93	11.18	4.84	11.69	9.95	10.63	12.9	13.07	**13.12**
Arm	4.11	3.35	3.9	4.92	12.11	5.65	12.68	22.58	20.54	**24.65**

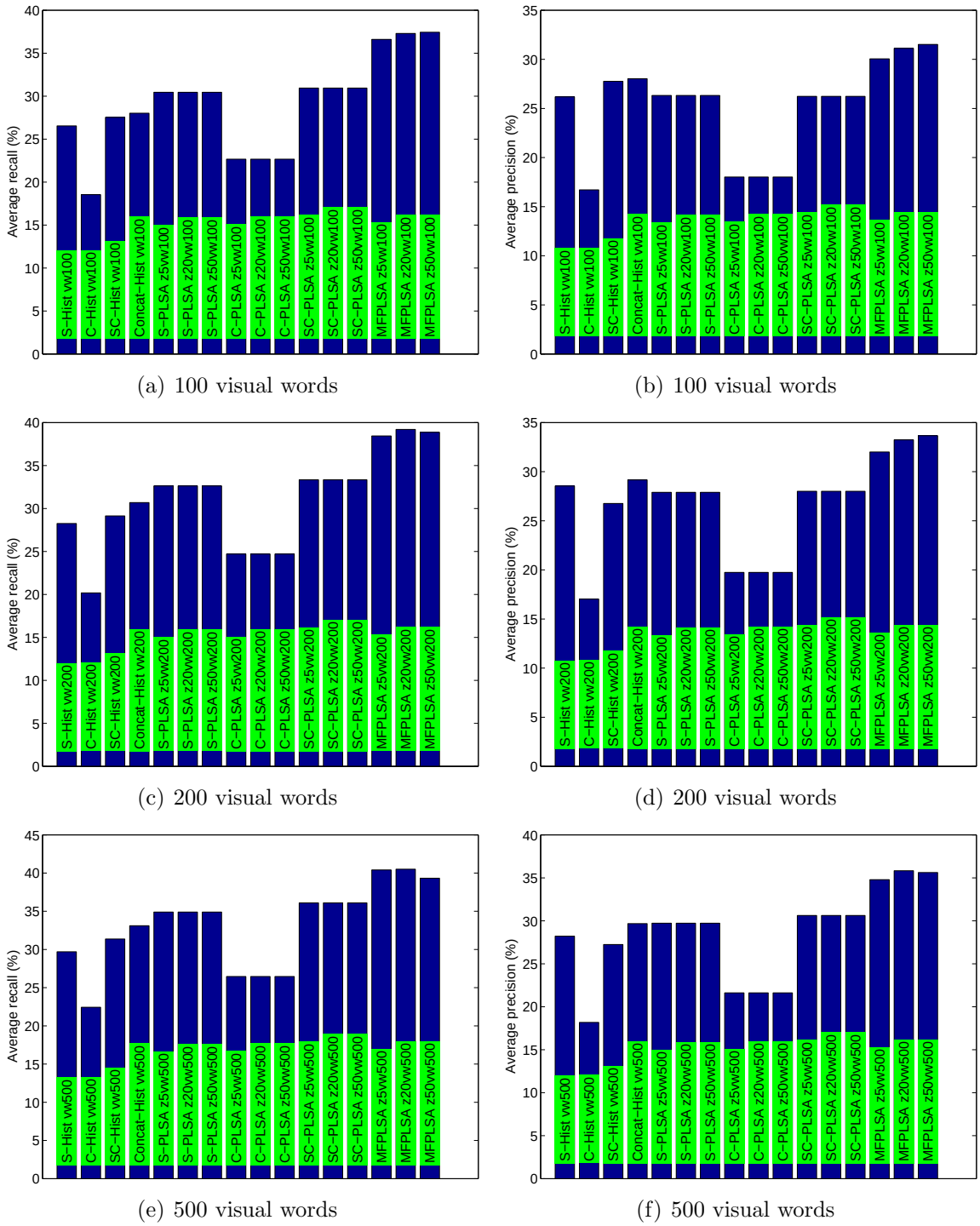


(a) 100 visual words (b) 100 visual words

(c) 200 visual words (d) 200 visual words

(e) 500 visual words (f) 500 visual words

Figure 5.8: The average recall and precision evaluated with the VOC2009 database. z denotes hidden topics, following which the number indicates the number of topics for pLSA-based approaches.



(a) 100 visual words (b) 100 visual words

(c) 200 visual words (d) 200 visual words

(e) 500 visual words (f) 500 visual words

Figure 5.9: The average recall and precision evaluated with the LabelMe database. z denotes hidden topics, following which the number indicates the number of topics for pLSA-based approaches.

as a mixture model to characterize $P_\omega(w)$, which is the distribution of visual words in each class. This means that each class has their own set of visual topics, which are the mixture components of $P_\omega(w)$. This learning and testing protocol keeps the pLSA-based approaches and the MF-pLSA different only in the aspect of utilizing features and identical otherwise. Therefore, the advantage resulting from the way MF-pLSA combines multiple features can be discovered. The underlying mechanism resulting in the little impact of changing the number of visual topics on the performance of pLSA-based approaches can be explained by taking three regions as examples shown in Fig. 5.10(a). It can be observed that, when the number of topics increases, the $P_\omega(w)$'s are decomposed into more topics while the mixture weights are decreased in that they constitute the distribution of topics over an image region. Furthermore, keeping the number of visual words equal, we consider the $P_\omega(w)$'s of each class estimated with different numbers of visual topics as vectors and calculate the mean of them. Then, we evaluate the difference between a $P_\omega(w)$ and its associated mean based on the L1-Norm. For C-pLSA, S-pLSA and SC-pLSA, the difference is calculated for all combinations of the numbers of visual words and topics. The result of the class of person from the LabelMe database is selected for illustration. It can be observed from Fig. 5.10(b) that the difference among the same type of pLSA-based models with the same number of visual words but different numbers of topics is nearly negligible.

To further compare the proposed MF-pLSA with other approaches, it is worthwhile to study the degree of the level of confusion across models of different concept classes. To this end, the confusion matrices of the models studied in our experiments with different experimental settings are shown in Fig. 5.11 through Fig. 5.15 for the VOC2009 database and in Fig. 5.16 and Fig. 5.20 for the LabelMe database. It can be observed from the results that histogram-based approaches can hardly learn discriminative

models for recognizing the object categories, even when SIFT and LTCH are combined through descriptor vector concatenation or BOVW vector concatenation. When pLSA-based approaches, i.e. S-pLSA, C-pLSA and SC-pLSA, are employed, the classification performance is improved thanks to the ability of pLSA to capture the co-occurrence pattern of data and avoid over-fitting. However, it can also be observed that the descriptor vector concatenation used with the pLSA-based models results in very limited performance improvement. To compare with the performance of MF-pLSA, the SC-pLSA is selected because, in terms of the $\mathcal{R}_{\omega}^{avg}$ and $\mathcal{P}_{\omega}^{avg}$, SC-pLSA generally has the second best performance after the MF-pLSA among all the approaches. As per the conventional definition of confusion matrix, each column of the matrix corresponds to a ground truth class and each row is associated with a predicted target class. Taking the comparison among Fig. 5.15(a), Fig. 5.15(d), Fig. 5.15(i) and Fig. 5.15(j) as an example, it can be observed that the values of many off-diagonal elements are reduced using the MF-pLSA. Using SC-pLSA, many samples of the class of person or potted plants are classified into the class of bottle, whereas the MF-pLSA effectively reduce the confusion between these classes. Moreover, the number of the regions of cars which are labeled as aeroplane using the SC-pLSA is also considerably reduced using the MF-pLSA. Although there are still a number of regions of boats that are categorized into the class of aeroplane, the recall of the boat class is still improved using MF-pLSA over SC-pLSA. It is also observed from the third column of Fig. 5.15 that, when the sizes of the SIFT and LTCH code books are 500, samples of different classes are more likely to be classified as person, especially the classes of bottle, cat, diningtable, dog and sofa. Nonetheless, further inspection on the confusion matrix reveals that the recall of the classes of cat, diningtable and dog obtained using MF-pLSA are actually higher than that obtained using SC-pLSA. Regarding LabelMe database, we find that the proposed feature fusion is capable of alleviating the

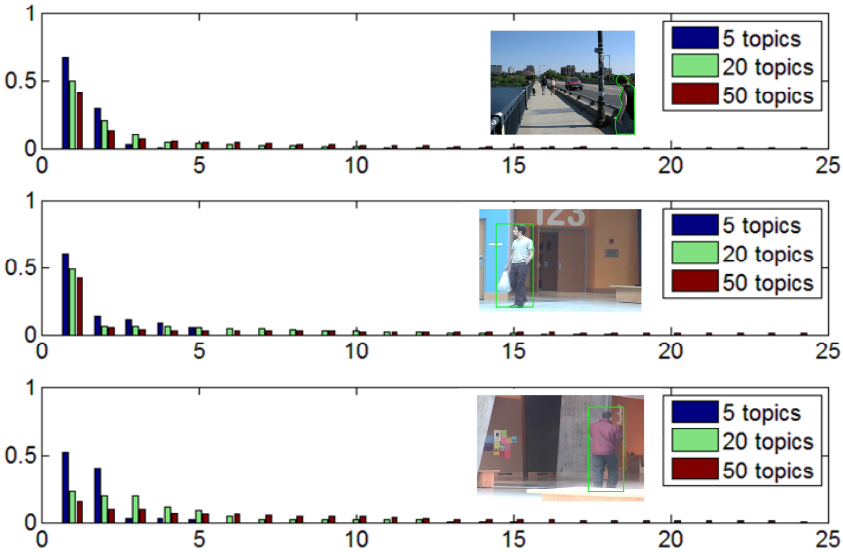
confusion between sidewalk and road. Misclassification, such as mistaking a door for a window, mistaking a car for road and mistaking a table for a car, is also reduced.

To visually inspect the comparison between the results of MF-pLSA and that of SC-pLSA, we select a number of samples from the classification results which are illustrated in Fig. 5.21 and Fig. 5.22. As shown in Fig. 5.21(a) and Fig. 5.21(g), the MF-pLSA correctly labels the objects of an aeroplane and a sofa although only part of the object of interest appears in the images. Shown in Fig. 5.21(c) and Fig. 5.21(d) are two different types of boats, where their scales and visual appearance are also quite different. Using MF-pLSA, however, both of the two objects of boat are successfully recognized. Another example of recognizing objects of the same class yet with different appearance can be observed in Fig. 5.22(g) and Fig. 5.22(h) for the class of building. In addition, the regions of partially occluded cars in Fig. 5.21(f) and Fig. 5.22(a) can also be successfully classified using the MF-pLSA.

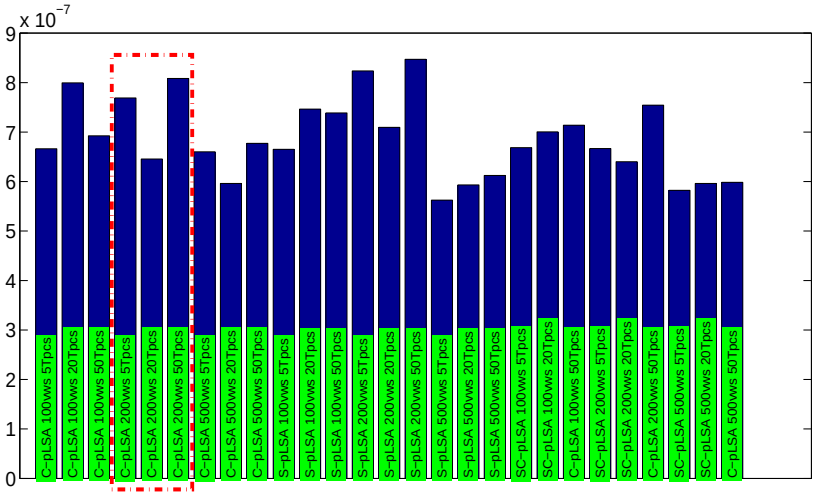
5.5 Summary

In this chapter, we present a novel framework of integrating texture and color descriptors which is applied to the problem of annotating object regions of images. Having the image data represented as 3-tuples extracted from key points, the new framework jointly learns the distributions of these two types of descriptors with a mixture model, termed MF-pLSA, for each of the object classes. The mixture models are used along with the supervised classification paradigm to classify previously unseen image regions into one of several pre-defined object categories. Compared with other approaches which combine these kinds of descriptors through descriptor level vector concatenation, the MF-pLSA only needs to learn the visual words in the individual domains of different descriptors

separately. Compared with the methods integrating descriptors via BOVW level vector concatenation, the distributions of different descriptors in MF-pLSA are parameterized independently but learned jointly. Moreover, the mixture distribution accounts for intra-class variation of the same object category. Through extensive experimental evaluation using the VOC2009 database and the LabelMe database as well as different experimental settings regarding the numbers of visual words and topics, the superiority of the MF-pLSA to seven other approaches is demonstrated, including descriptor level vector concatenation, i.e. SC-Hist and SC-pLSA, and BOVW level vector concatenation, i.e. Concat-Hist. Compared with the second best approach, i.e. SC-pLSA, the MF-pLSA brings about the performance improvement of up to 3% and 6.5% in terms of the average recall, using the VOC2009 and LabelMe databases. As far as the average precision is considered, the performance is improved by 2.0% and 5.2% over the SC-pLSA using the MF-pLSA with the VOC2009 and LabelMe databases. By using the confusion matrices, it can be observed, from the reduced values of the off-diagonal elements, that the MF-pLSA enhances the discriminative powers of the statistical models of individual object categories. There are still some object categories, for which the MF-pLSA does not lead to the best performance compared with the others in our experiments. We believe introducing contextual information, which encodes the co-occurrence of different object categories, or a hierarchical organization of them into the presented framework should be worth exploring and potentially able to further improve the performance.

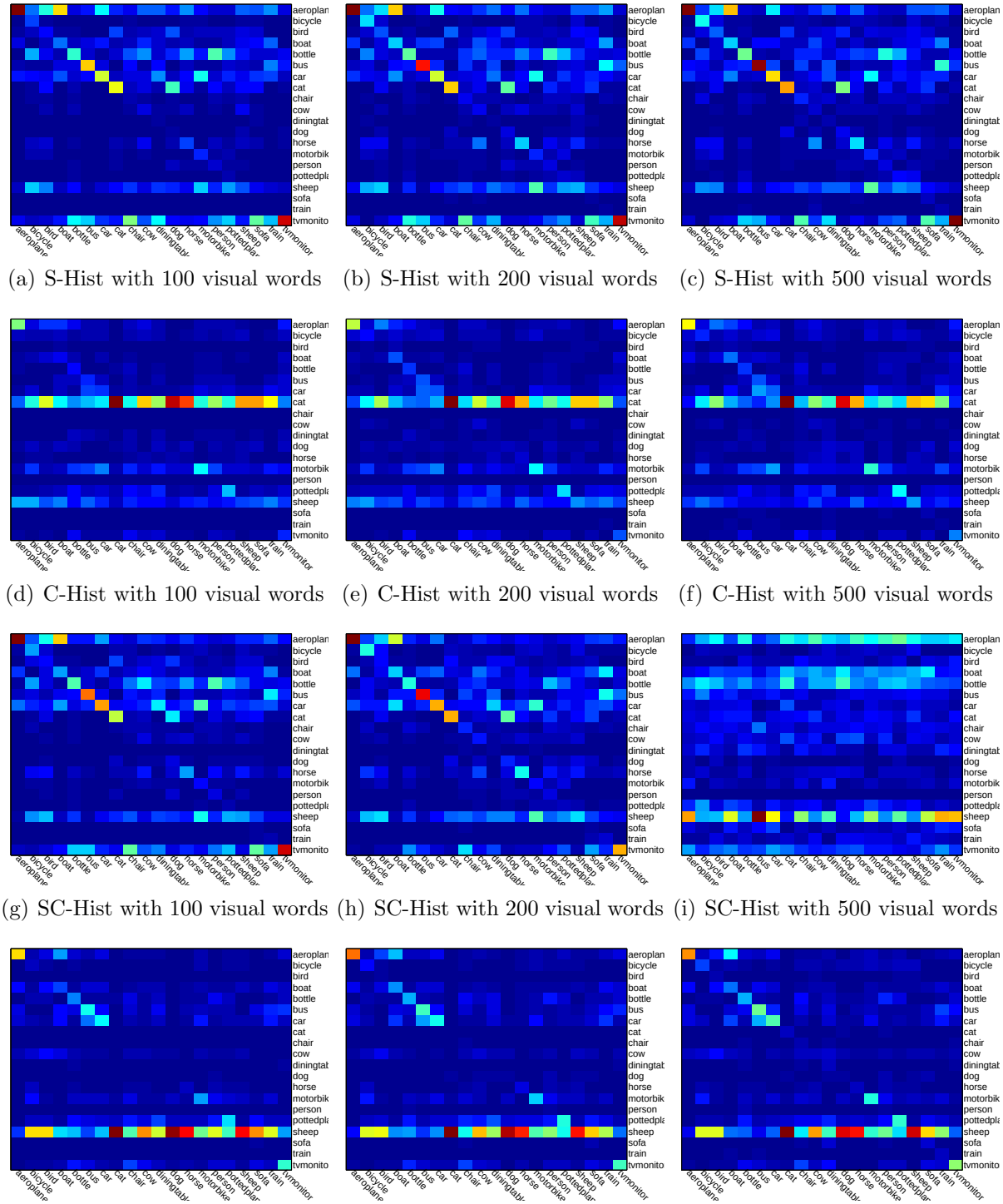


(a)



(b)

Figure 5.10: The variation of topic mixture $P(z|d)$ after changing the number of visual topics. The class chosen for this illustration is person from the LabelMe database. (a) The topic mixture of three sample regions. For each region, the $P(z|d)$ are sorted in descending order based on the values. (b) The comparison among the pLSA-based models with different numbers of topics. To see the little impact of changing the number of topics on the resulting pLSA, select a kind of visual word, e.g. LTCH, and a certain number of visual words, e.g. 200, as illustrated with the red rectangle in red dashed line.



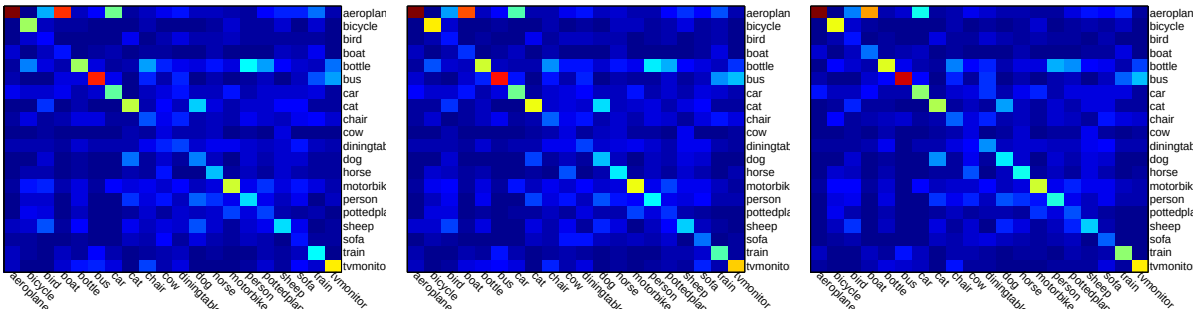
(a) S-Hist with 100 visual words (b) S-Hist with 200 visual words (c) S-Hist with 500 visual words

(d) C-Hist with 100 visual words (e) C-Hist with 200 visual words (f) C-Hist with 500 visual words

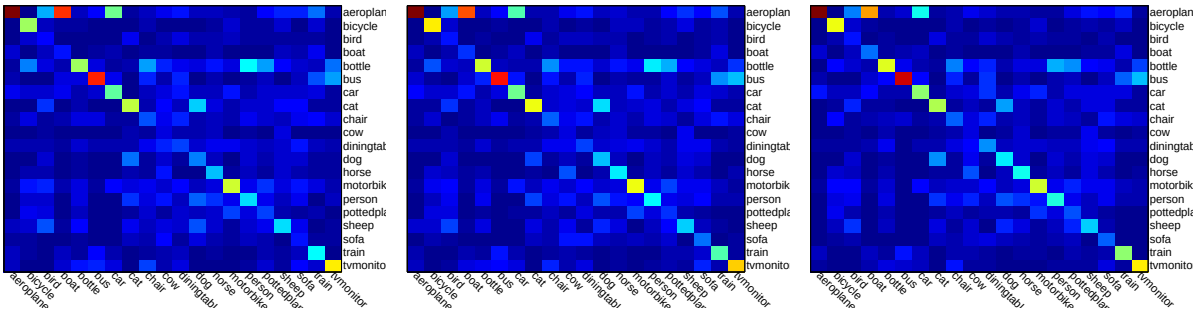
(g) SC-Hist with 100 visual words (h) SC-Hist with 200 visual words (i) SC-Hist with 500 visual words

(j) Concat-Hist with 100 visual words (k) Concat-Hist with 200 visual words (l) Concat-Hist with 500 visual words

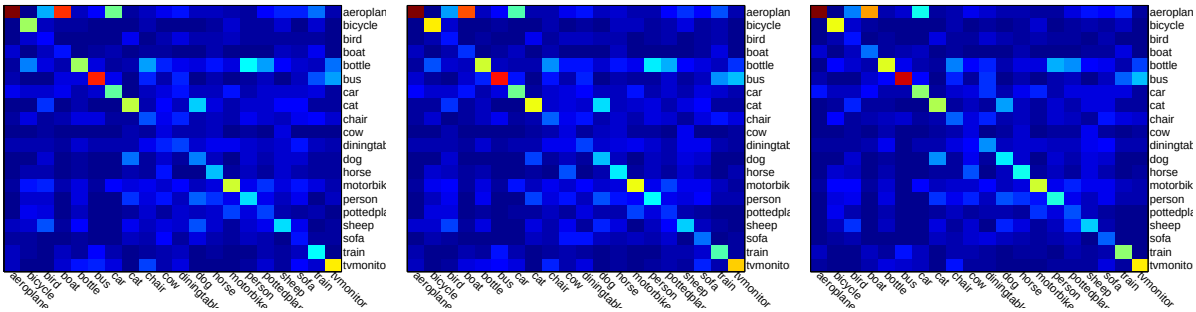
Figure 5.11: Confusion matrices of the S-Hist, C-Hist, SC-Hist and Concat-Hist evaluated using the VOC2009 database. Each row includes the results of one of the four approaches.



(a) S-pLSA with 100 visual words with 5 topics (b) S-pLSA with 200 visual words with 5 topics (c) S-pLSA with 500 visual words with 5 topics

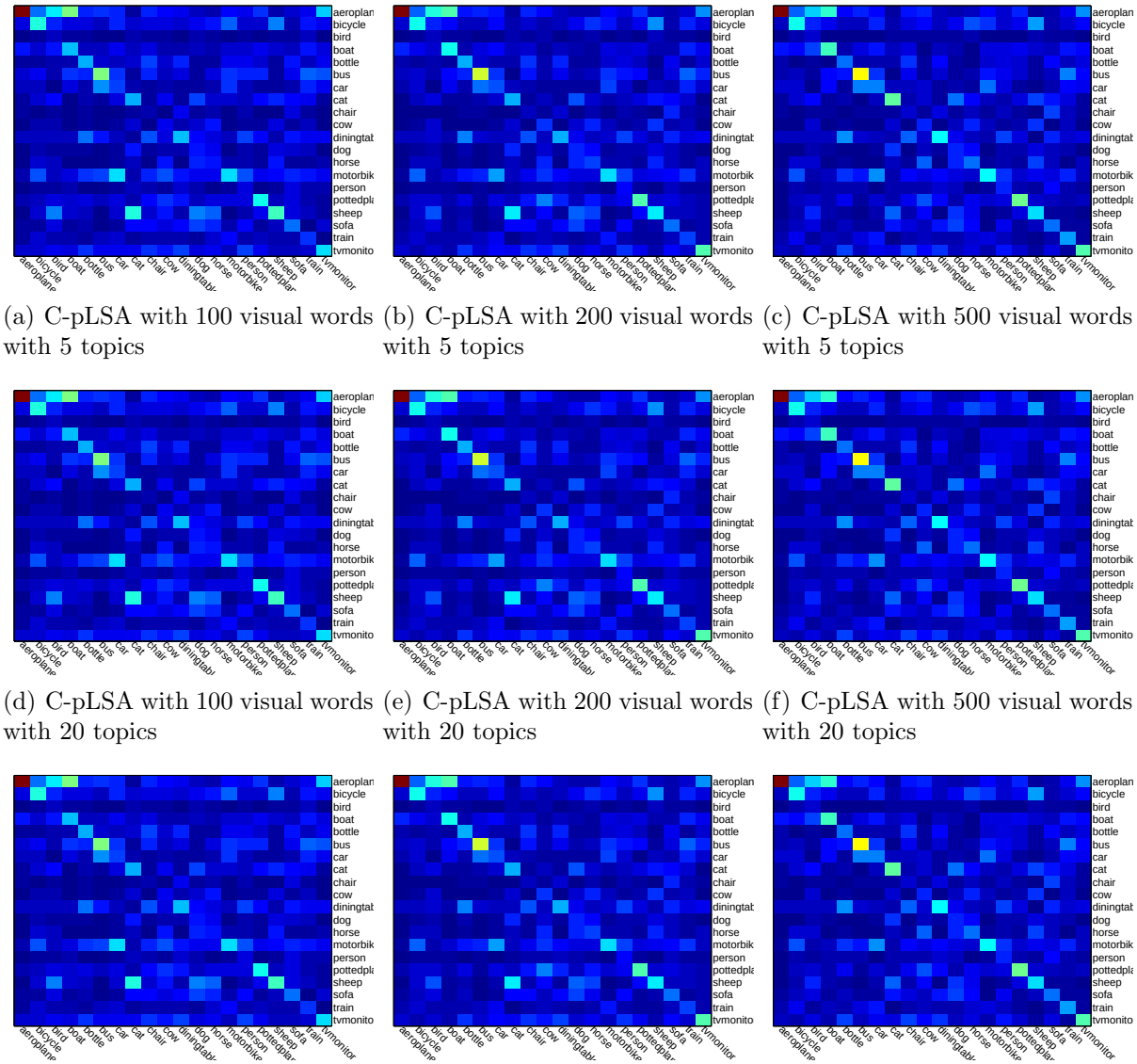


(d) S-pLSA with 100 visual words with 20 topics (e) S-pLSA with 200 visual words with 20 topics (f) S-pLSA with 500 visual words with 20 topics



(g) S-pLSA with 100 visual words with 50 topics (h) S-pLSA with 200 visual words with 50 topics (i) S-pLSA with 500 visual words with 50 topics

Figure 5.12: Confusion matrices of the S-pLSA with different numbers of visual words and visual topics. evaluated using the VOC2009 database.



(a) C-pLSA with 100 visual words with 5 topics

(b) C-pLSA with 200 visual words with 5 topics

(c) C-pLSA with 500 visual words with 5 topics

(d) C-pLSA with 100 visual words with 20 topics

(e) C-pLSA with 200 visual words with 20 topics

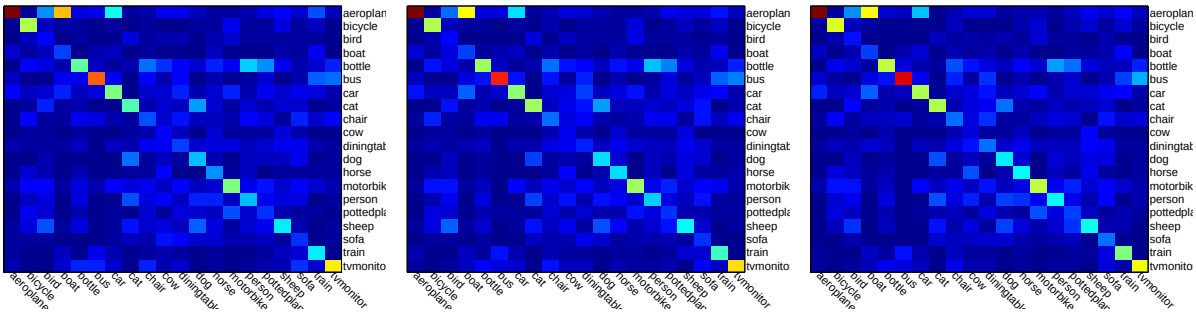
(f) C-pLSA with 500 visual words with 20 topics

(g) C-pLSA with 100 visual words with 50 topics

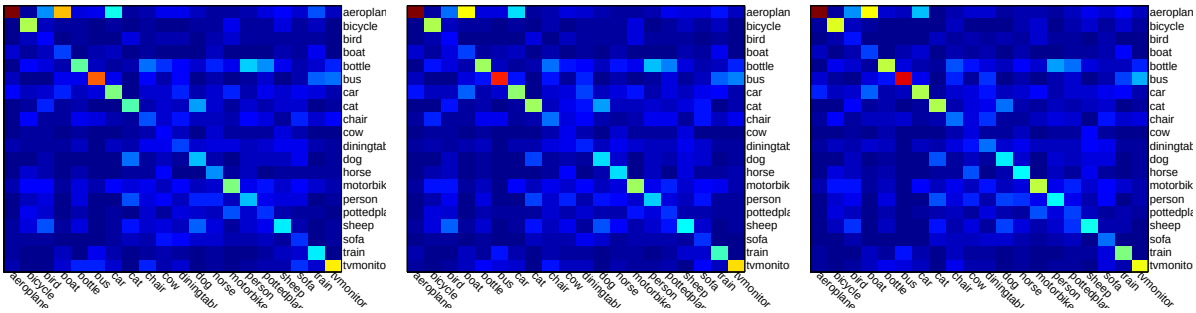
(h) C-pLSA with 200 visual words with 50 topics

(i) C-pLSA with 500 visual words with 50 topics

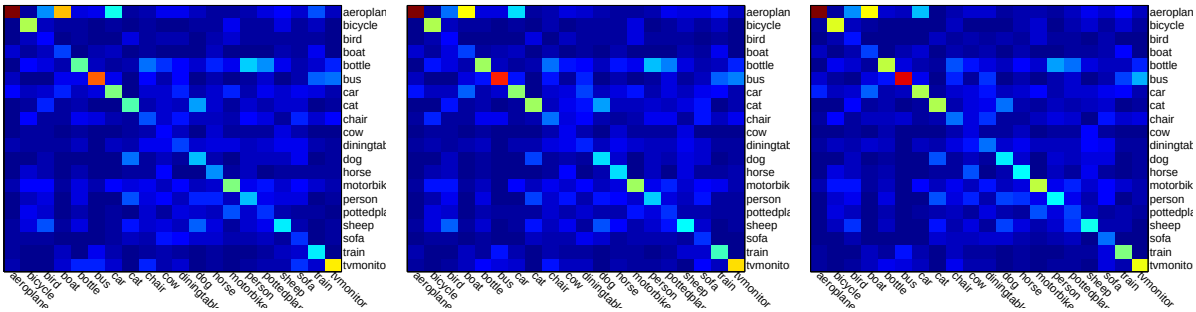
Figure 5.13: Confusion matrices of the C-pLSA with different numbers of visual words and visual topics evaluated using the VOC2009 database.



(a) SC-pLSA with 100 visual words with 5 topics (b) SC-pLSA with 200 visual words with 5 topics (c) SC-pLSA with 500 visual words with 5 topics

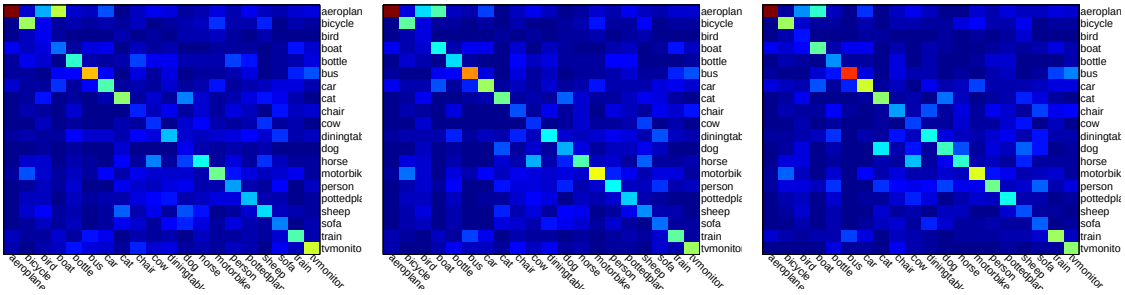


(d) SC-pLSA with 100 visual words with 20 topics (e) SC-pLSA with 200 visual words with 20 topics (f) SC-pLSA with 500 visual words with 20 topics

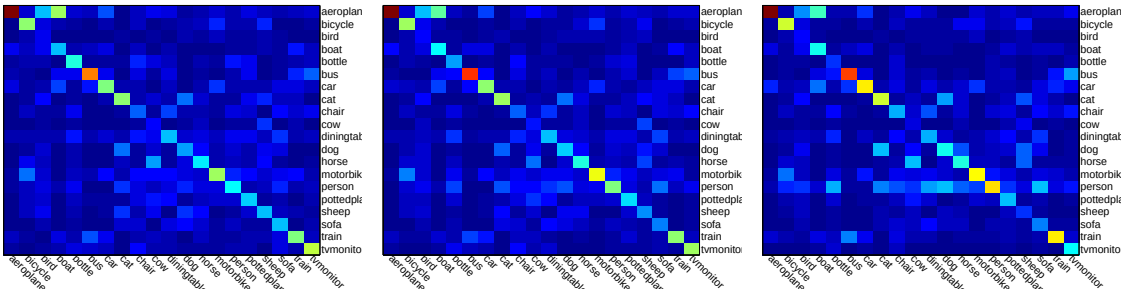


(g) SC-pLSA with 100 visual words with 50 topics (h) SC-pLSA with 200 visual words with 50 topics (i) SC-pLSA with 500 visual words with 50 topics

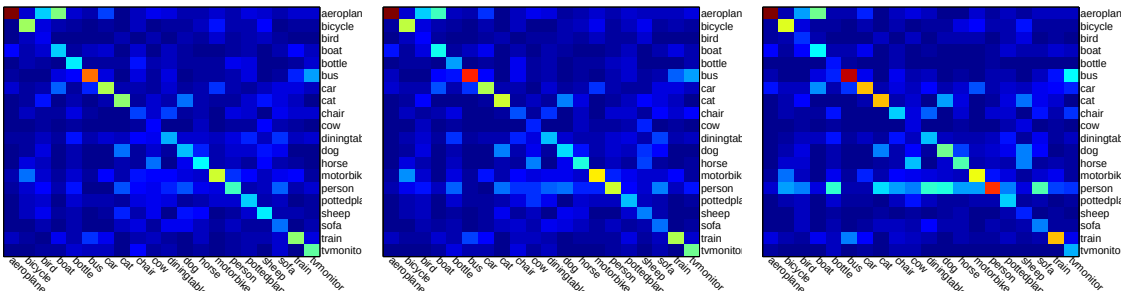
Figure 5.14: Confusion matrices of the SC-pLSA with different numbers of visual words and visual topics evaluated using the VOC2009 database.



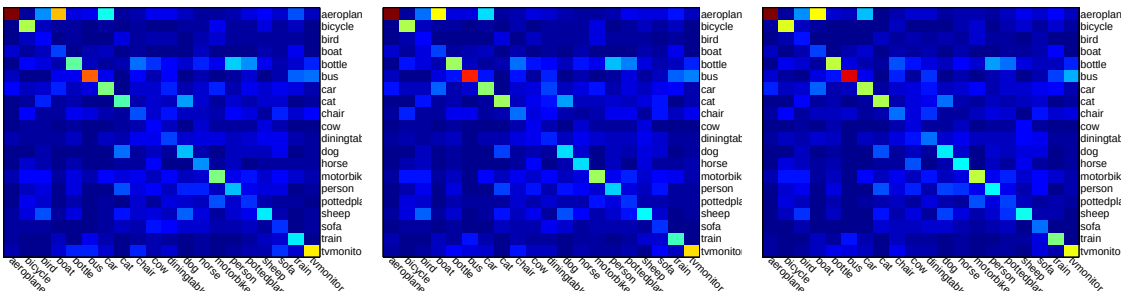
(a) MF-pLSA: 100 visual words and 5 visual topics (b) MF-pLSA: 200 visual words and 5 visual topics (c) MF-pLSA: 500 visual words and 5 visual topics



(d) MF-pLSA: 100 visual words and 20 visual topics (e) MF-pLSA: 200 visual words and 20 visual topics (f) MF-pLSA: 500 visual words and 20 visual topics

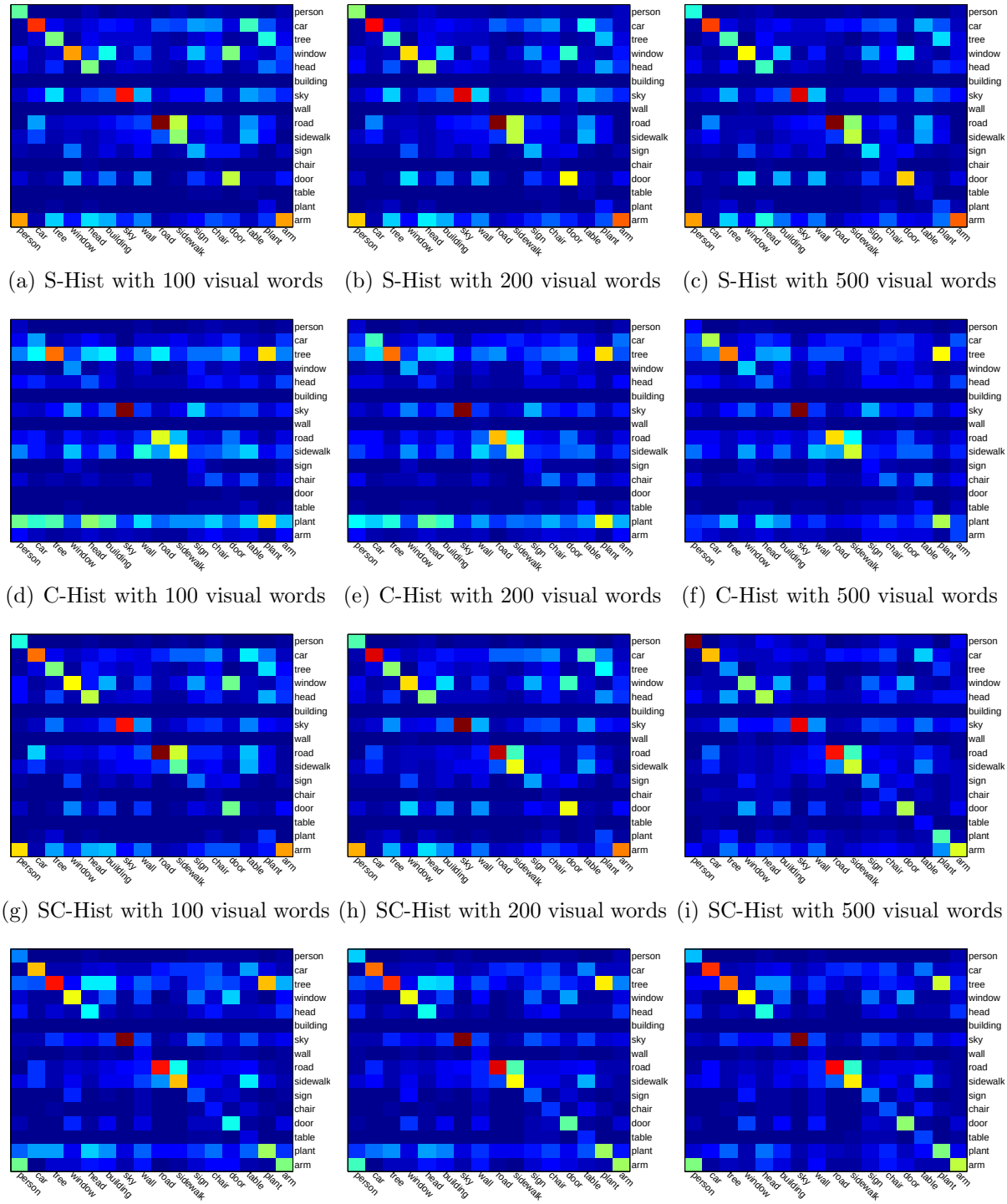


(g) MF-pLSA: 100 visual words and 50 visual topics (h) MF-pLSA: 200 visual words and 50 visual topics (i) MF-pLSA: 500 visual words and 50 visual topics



(j) SC-pLSA: 100 visual words and 5 visual topics (k) SC-pLSA: 200 visual words and 5 visual topics (l) SC-pLSA: 500 visual words and 5 visual topics

Figure 5.15: Confusion matrices of the MF-pLSA and SC-pLSA evaluated using the VOC2009 database. The comparison should be performed across the figures along each column, which includes the results obtained with the same number of visual words. For the SC-pLSA, only the results obtained with 5 visual topics are included since the other numbers of visual words lead to the same classification results as discussed before.



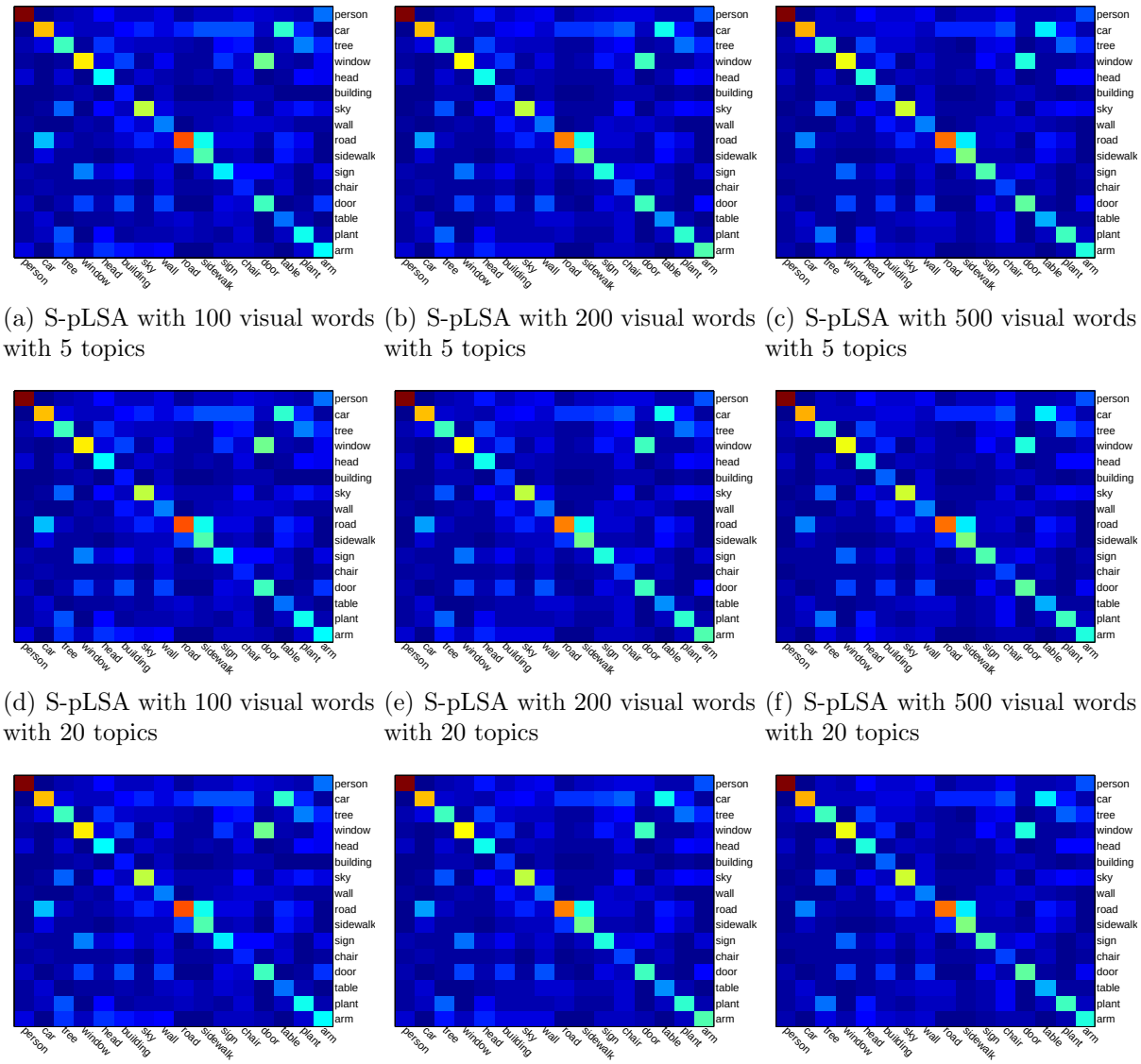
(a) S-Hist with 100 visual words (b) S-Hist with 200 visual words (c) S-Hist with 500 visual words

(d) C-Hist with 100 visual words (e) C-Hist with 200 visual words (f) C-Hist with 500 visual words

(g) SC-Hist with 100 visual words (h) SC-Hist with 200 visual words (i) SC-Hist with 500 visual words

(j) Concat-Hist with 100 visual words (k) Concat-Hist with 200 visual words (l) Concat-Hist with 500 visual words

Figure 5.16: Confusion matrices of the S-Hist, C-Hist, SC-Hist and Concat-Hist evaluated using the LabelMe database. Each row includes the results of one of the four approaches.



(a) S-pLSA with 100 visual words with 5 topics

(b) S-pLSA with 200 visual words with 5 topics

(c) S-pLSA with 500 visual words with 5 topics

(d) S-pLSA with 100 visual words with 20 topics

(e) S-pLSA with 200 visual words with 20 topics

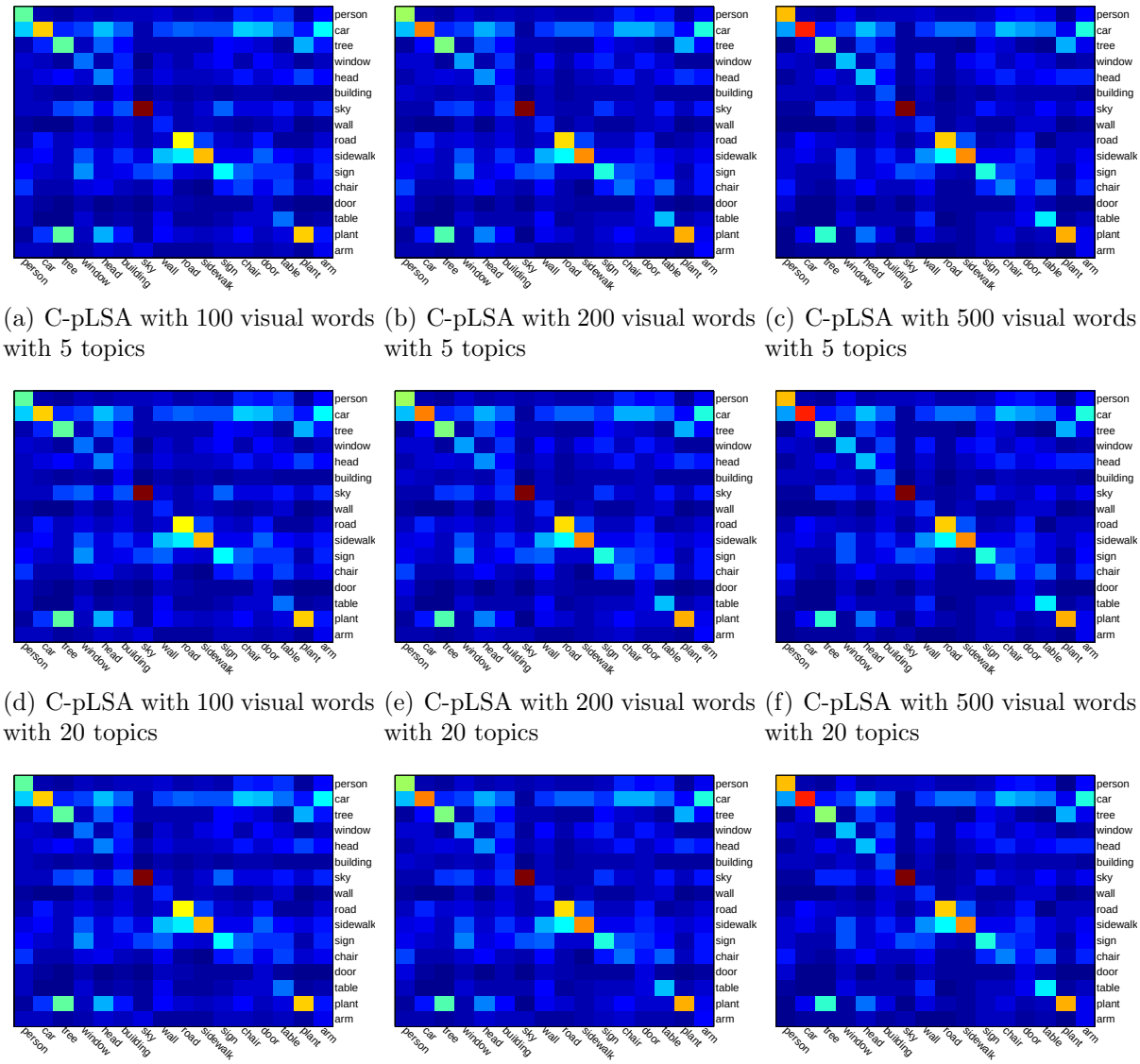
(f) S-pLSA with 500 visual words with 20 topics

(g) S-pLSA with 100 visual words with 50 topics

(h) S-pLSA with 200 visual words with 50 topics

(i) S-pLSA with 500 visual words with 50 topics

Figure 5.17: Confusion matrices of the S-pLSA with different numbers of visual words and visual topics. evaluated using the LabelMe database.

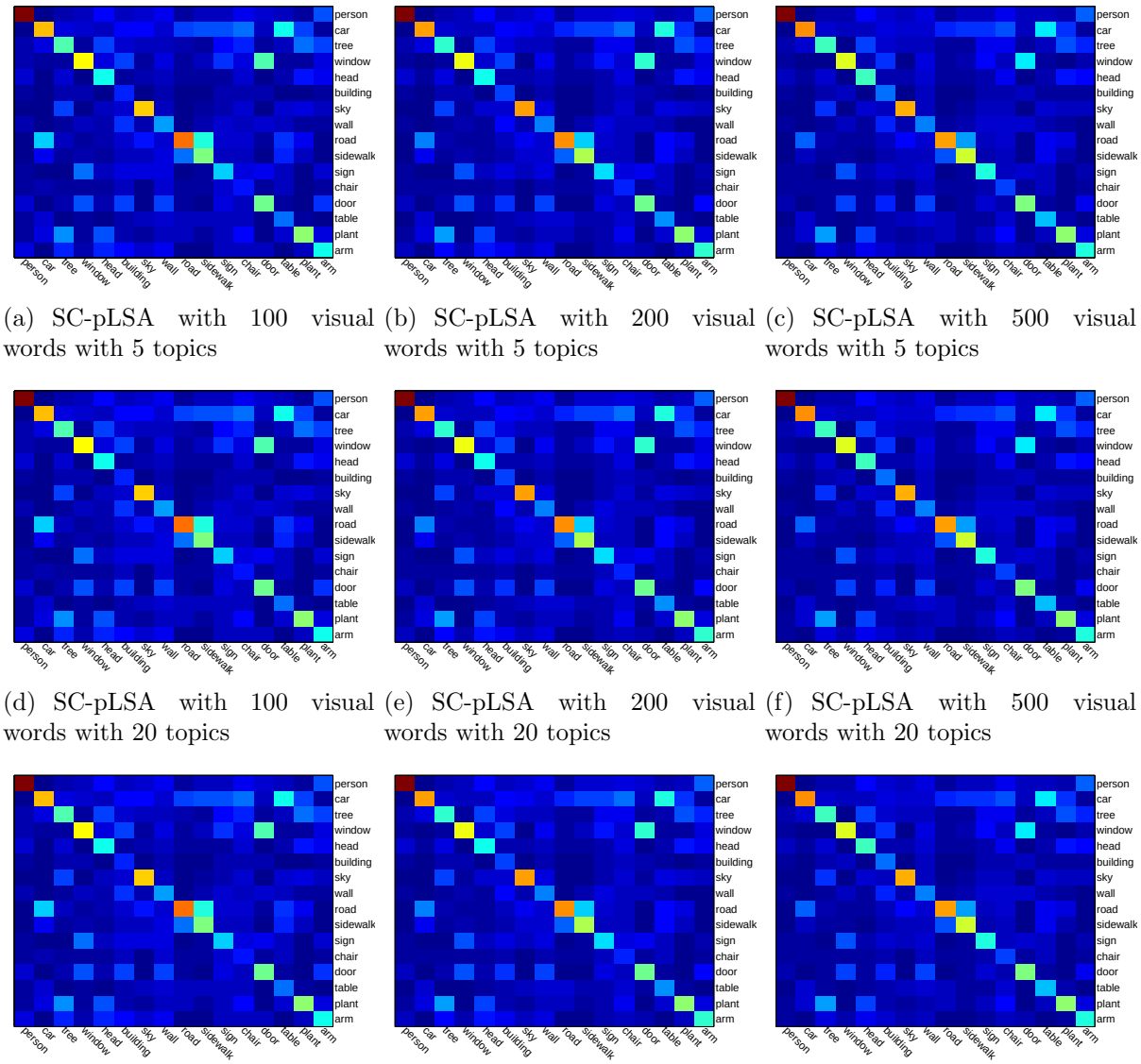


(a) C-pLSA with 100 visual words with 5 topics (b) C-pLSA with 200 visual words with 5 topics (c) C-pLSA with 500 visual words with 5 topics

(d) C-pLSA with 100 visual words with 20 topics (e) C-pLSA with 200 visual words with 20 topics (f) C-pLSA with 500 visual words with 20 topics

(g) C-pLSA with 100 visual words with 50 topics (h) C-pLSA with 200 visual words with 50 topics (i) C-pLSA with 500 visual words with 50 topics

Figure 5.18: Confusion matrices of the C-pLSA with different numbers of visual words and visual topics evaluated using the LabelMe database.



(a) SC-pLSA with 100 visual words with 5 topics

(b) SC-pLSA with 200 visual words with 5 topics

(c) SC-pLSA with 500 visual words with 5 topics

(d) SC-pLSA with 100 visual words with 20 topics

(e) SC-pLSA with 200 visual words with 20 topics

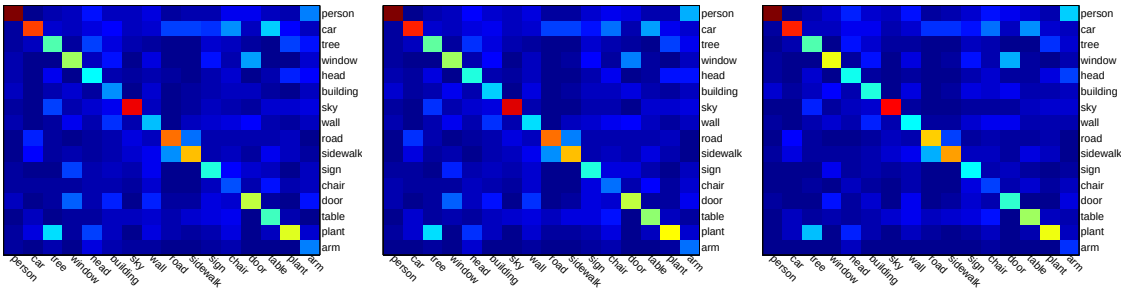
(f) SC-pLSA with 500 visual words with 20 topics

(g) SC-pLSA with 100 visual words with 50 topics

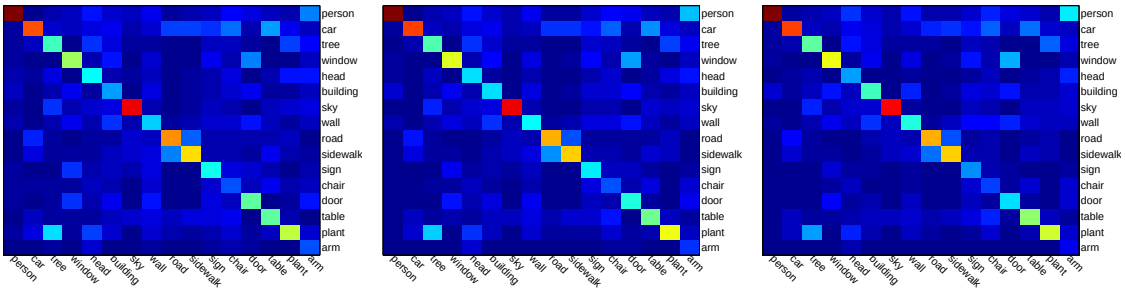
(h) SC-pLSA with 200 visual words with 50 topics

(i) SC-pLSA with 500 visual words with 50 topics

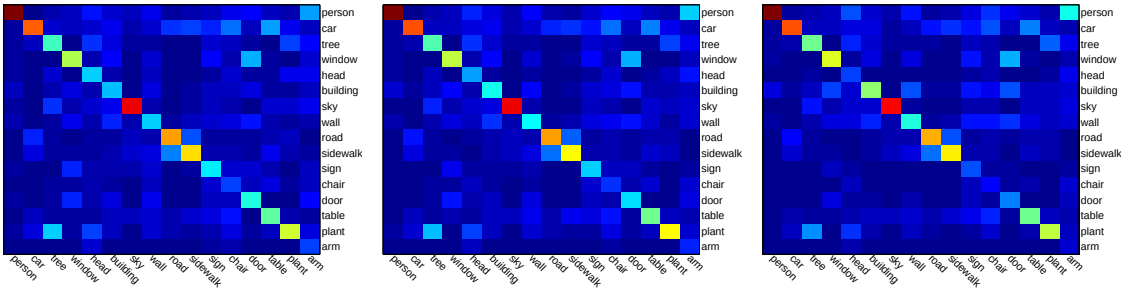
Figure 5.19: Confusion matrices of the SC-pLSA with different numbers of visual words and visual topics evaluated using the LabelMe database.



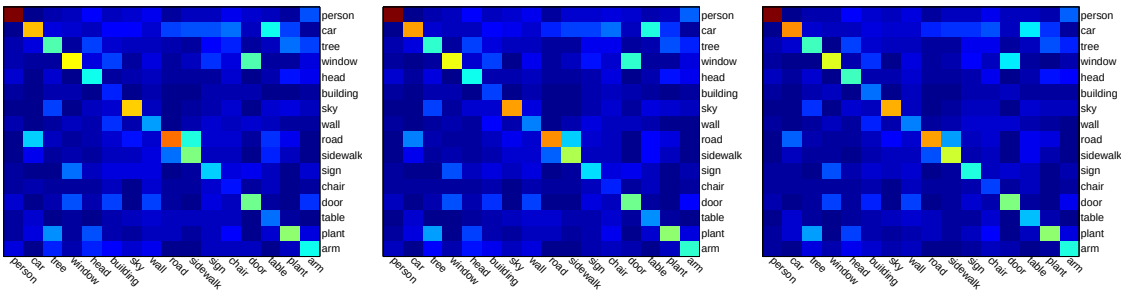
(a) MF-pLSA: 100 visual words and 5 visual topics (b) MF-pLSA: 200 visual words and 5 visual topics (c) MF-pLSA: 500 visual words and 5 visual topics



(d) MF-pLSA: 100 visual words and 20 visual topics (e) MF-pLSA: 200 visual words and 20 visual topics (f) MF-pLSA: 500 visual words and 20 visual topics

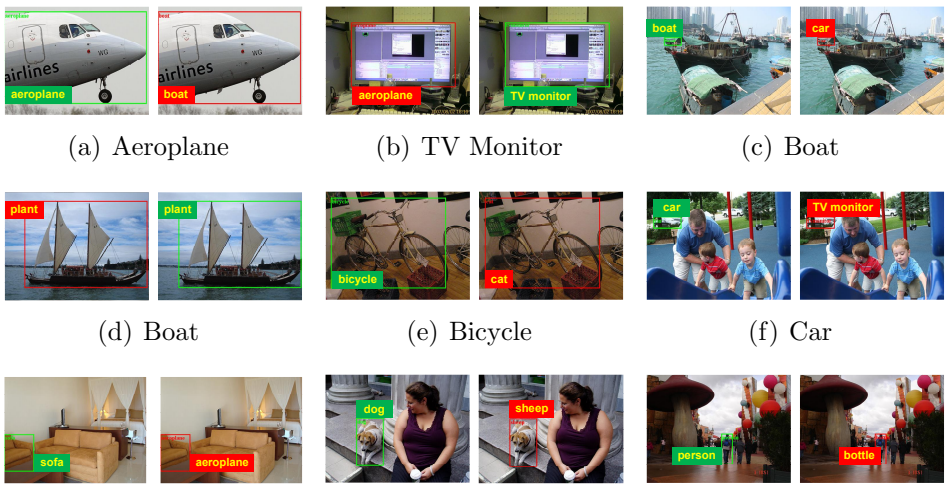


(g) MF-pLSA: 100 visual words and 50 visual topics (h) MF-pLSA: 200 visual words and 50 visual topics (i) MF-pLSA: 500 visual words and 50 visual topics



(j) SC-pLSA: 100 visual words and 5 visual topics (k) SC-pLSA: 200 visual words and 5 visual topics (l) SC-pLSA: 500 visual words and 5 visual topics

Figure 5.20: Confusion matrices of the MF-pLSA and SC-pLSA evaluated using the LabelMe database. The comparison should be performed across the figures along each column, which includes the results obtained with the same number of visual words. For the SC-pLSA, only the results obtained with 5 visual topics are included since the other numbers of visual words lead to the same classification results as discussed before.

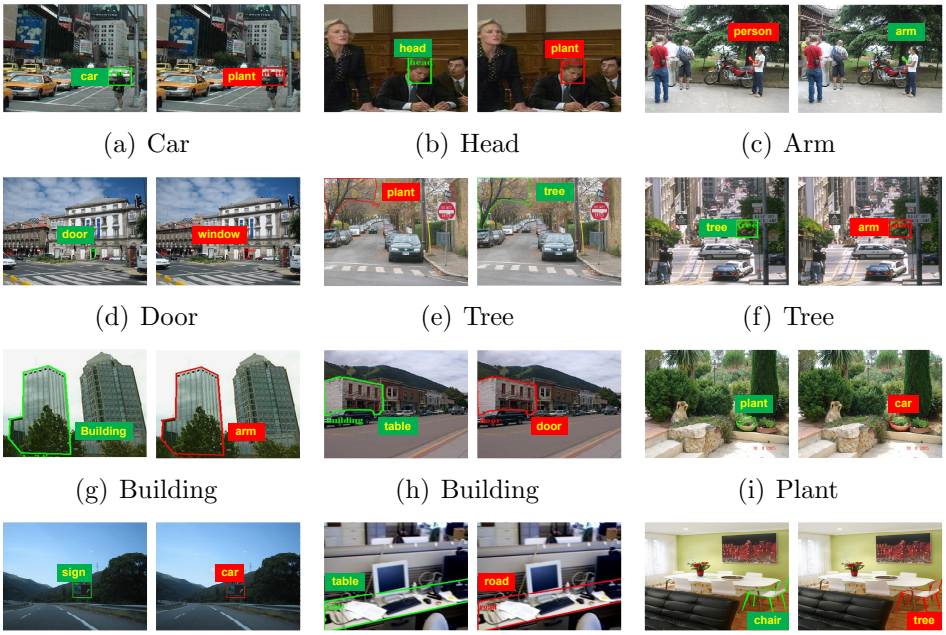


(a) Aeroplane (b) TV Monitor (c) Boat

(d) Boat (e) Bicycle (f) Car

(g) Sofa (h) Dog (i) Person

Figure 5.21: Examples of classification results using the VOC2009 database. The figure on the left-hand side of each class is the result of MF-pLSA and the one on the right-hand side is the result of SC-pLSA. In addition, the boundary of a successfully classified object is shown in green.



(a) Car (b) Head (c) Arm

(d) Door (e) Tree (f) Tree

(g) Building (h) Building (i) Plant

(j) Sign (k) Table (l) Chair

Figure 5.22: Examples of classification results using the LabelMe database. The figure on the left-hand side of each class is the result of MF-pLSA and the one on the right-hand side is the result of SC-pLSA. In addition, the boundary of a successfully classified object is shown in green.

Chapter 6

Conclusions and Future Work

6.1 Summary

This thesis is primarily focused on the task of integrating available information sources at different levels of the pattern classification for the applications of image annotation and retrieval. To be specific, there are two levels considered in the presented works. Based on the order of information processing within a general pattern classification system, these two levels are:

- Within the low-level visual domain, where various visual features are available for representing the image information. Common characteristics of these features are that each type of features is image-specific and their extraction from a single image is a relatively short-term process. On the other hand, the difference among them lies in the fact that they describe distinct aspects of the visual properties of image data.

- Across the low-level visual and high-level contextual domains, where the available

pieces of information are heterogeneous compared with the relation of the low-level visual features. Rather than being image specific, the contextual information encodes the dependence across different semantic concepts or images, which can be considered as a database-wide feature of image data. In addition, the acquisition of the contextual information by means of machine learning is a long-term process in contrast to the short-term feature extraction in the visual domain.

- Between the audio and visual domains, considering that there is strong correlation between the visual properties of and the sound made certain types of objects. This can be considered as the same level as the second one.

Based on the above analysis, two different approaches respectively targeting the two different levels are proposed and implemented.

In Chapter 3, a method for the joint exploitation of the low-level visual features and the high-level contextual information is proposed. Considering the different nature of these two types of information and their distinct roles in the pattern classification process, a general Bayesian framework is proposed, which consists of a content and a context component corresponding to the two different domains. Rather than jointly modeling the statistical distribution of the visual features and context, the Bayesian framework utilizes the former for likelihood evaluation and the latter for *a priori* probability evaluation. Within such a framework of information exploitation, the content and context components refine each others' results in terms of the degrees of relevance of the visual feature of a to-be-annotated region with respect to a semantic class.

The first application of the Bayesian framework is image annotation, resulting in the CBIA framework dealing with the visual recognition task that is region-based, meaning each region is given a specific description upon the semantics. In the case of our

study, such description is provided in the form of a keyword, representing a concrete real world object category. To reduce UMI and make the annotation fully automatic, a content-based search module was added to the annotation framework. The search module is responsible for suggesting relevant keywords which are in turn used to infer the probabilities of other semantic concepts with the context component. Through extensive experiments with various datasets, the advantage of the CBIA framework was demonstrated by its performance superior to those content-based and context-based approaches. However, the CBIA also has limitations. For example, the statistical dependence across different to-be-annotated regions is not explicitly considered in the CBIA framework. In this sense, the contextual information utilized in the CBIA framework is similar to the gist context. Meanwhile, they are also different in that the gist is based on global visual statistics whereas the contextual information of the CBIA is based on semantic concepts.

The second application of the Bayesian framework is image retrieval, resulting in the CLBIR framework, where the primary consideration is how to leverage the knowledge learned within each retrieval session as well as across multiple retrieval sessions. To facilitate the acquisition of such knowledge, the content and context are learned through STRF and LTRF. The CLBIR framework has the flexibility that the content component can be instantiated with various types of approaches to CBIR, i.e. their distance functions for similarity measure. In our experiments, this flexibility was demonstrated using a L1-Norm and an SVMAL content component respectively. Meanwhile, the combination of content and context of the Bayesian framework enables the CLBIR to work even under the circumstance of no available high-level contextual information. With the ability of gradually accumulating the past retrieval results and incrementally updating the context component, the CLBIR framework has the functionality of memorizing learned knowledge through a long-term process. At the same time, each user still can use the STRF to

polish the query formulation based on his/her own information need or preferences. On the other hand, the CLBIR framework still lacks a learning procedure which is capable of incorporating new images into the database automatically. A possible solution to this problem is to reformulate the contextual information as the statistical dependence across different semantic clusters within the database so that new images can be included without expanding the context model unless a new cluster has to be generated.

Moreover, the Bayesian framework has also been exploited to integrate the information from the audio and visual domains. In this application, considering the importance of the temporal information of audio sequences, HMMs are used to model the feature distribution within the audio domain, rather than the maximum entropy approach explored in the second chapter. In addition, non-parametric classification and adapted L1-norm are employed in the visual domain to propagate class-dependent audio relevance to the candidate images and measure the similarity between candidate images and a query, respectively. Along with the audio and visual relevance feedback, the performance of multi-modal framework is demonstrated superior to the image retrieval employing only the visual features.

In Chapter 5, the integration of multiple low-level features in the visual domain is addressed. The major line of thought is to learn the joint distribution of different visual features. The underlying principle is the more discriminative and informative frequent patterns induced from multiple visual properties. To this end, a model named MF-pLSA is proposed and is exploited to integrate two kinds of low-level visual descriptors, i.e. SIFT and LTCH. A supervised classification algorithm using the MF-pLSA is also developed. As mentioned earlier, if an image is considered as a function whose domain and range correspond to pixel location and intensities of different color channels, these two features essentially based on the function itself and its derivative. Based on the

BOVW image representation, the MF-pLSA characterizes the joint distribution of SIFT and LTCH visual words using a mixture model of discrete data. While being able to distinguish different semantic classes, the MF-pLSA can also discover different visual topics based on multiple features within each semantic class, which offers the opportunity to discovering various visual patterns among the objects of the same category. Extensive experiments were conducted, evaluating the performance of the MF-pLSA based on many different parameter settings, including the numbers of visual word and visual topics. Comparison was drawn between MF-pLSA and many different approaches, including those based on a single visual feature and vector concatenation of low level features. In addition, MF-pLSA was also compared with other machine learning approaches, such as nearest neighbor and pLSA for both single feature and multiple features. The superior performance of MF-pLSA was demonstrated through the criteria of recall and precision.

6.2 Future Work

Despite the intensive research effort within the past two decades, image retrieval and annotation are still quite challenging because of the difficulty lying in the semantic gap and sensory gap. Essentially, from the view points of machine learning and pattern classification, the problems associated with image retrieval and annotation yet to be solved are the same as those related to visual recognition. In terms of the future research work, the following directions are worth further exploring.

- There are still problems associated with the combination of information sources for image annotation and retrieval. For example, a unified learning and classification framework aiming at information combination at different levels simultaneously is desirable.

- A picture is worth a thousand words. This implies the rich semantics of images and results in the fact that the number of semantic categories involved in the visual recognition for image annotation and retrieval is very large. A large number of semantic categories leads to more severe semantic gap. Therefore, the scalability of statistical models in terms of handling a large number of semantic classes should be addressed properly.

- The resource on the Internet, such as images, text and their associated hyper-links, is very useful for multimedia data mining. Machine learning leveraging the web-scale databases are faced with the scalability issue as well. In this situation, parallel computing has huge potential to accelerating the computational procedure.

Appendix A

Publications

1. R. Zhang, L. Guan, "Integrating SIFT and Color Descriptor for Image Annotation", *submitted to IEEE Transaction on Image Processing.*

2. R. Zhang, L. Zhang, X.-J. Wang, L. Guan, "Multi-Feature pLSA for Combining Visual Features in Image Annotation," *ACM Multimedia*, 2011.

3. R. Zhang, L. Guan, "Collaborative Bayesian Image Annotation and Retrieval," *Machine Learning Techniques for Adaptive Multimedia Retrieval: Technologies Applications and Perspectives*, IGI Global, 2011, pp. 146-169.

4. L. Guan, Y. Wang, R. Zhang, Y. Tie, A. Bulzacki, M. T. Ibrahim, "Multimodal Information Fusion for Selected Multimedia Applications," *International Journal on Multimedia Intelligence and Security*, Vol. 1, No. 1, 2010, pp. 5-32.

5. R. Zhang, L. Guan, "A Bayesian Image Retrieval Framework," *International Journal of Digital Library Systems (IJDLS)*, Vol. 1, No. 2, 2010, pp. 43-58.

6. Y. Wang, R. Zhang, L. Guan, A. N. Venetsanopoulos, "Kernel fusion of audio

and visual information for emotion recognition", *in Proc. ICIAR*, Burnaby, BC, Canada, June 22-24, 2011, pp. 140-150.

7. R. Zhang, L. Guan, "A Bayesian Image Annotation Framework Integrating Search and Context," *in Proc. IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2010, pp. 499-504.

8. R. Zhang, M. T. Ibrahim, L. Guan, "A Collaborative Bayesian Image Retrieval Framework," *in Proc. Visual Communications and Image Processing (VCIP)*, 2010. [Prototype System Demo]

9. R. Zhang, L. Guan, "Multimodal Image Retrieval via Bayesian Information Fusion," *in Proc. IEEE International Conference on Multimedia and Expo (ICME)*, 2009, pp. 830-833.

10. L. Guan, P. Muneesawang, Y. Wang, R. Zhang, Y. Tie, A. Bulzacki, M. T. Ibrahim, "Multimedia Multimodal Methodologies," *in Proc. IEEE Conference on Multimedia and Expo (ICME)*, 2009, pp. 1600-1603.

11. R. Zhang, L. Guan, "A Collaborative Bayesian Image Retrieval Framework," *in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009, pp. 1953-1956. (Live demo available at: http://clbir.rml.ryerson.ca/main.htm)

12. R. Zhang, K. Wu, K.-H. Yap, L. Guan, "A Collaborative Bayesian Framework of Image Annotation," *in Proc. of Pacific-Rim Conference on Multimedia (PCM)*, 2008, pp. 348-357.

13. R. Zhang, L. Guan, "A New Framework of Relevance Feedback of Content-free Image Retrieval," *in Proc. of IEEE International Workshop on Multimedia Signal*

Processing (MMSP), 2008, pp. 685-690.

14. R. Zhang, X.P. Zhang, L. Guan, "Wavelet-based Texture Retrieval Using Independent Component Analysis," *in Proc. of IEEE International Conference on Image Processing (ICIP)*, 2007, pp. 341-344.

15. R. Zhang, X.P. Zhang, K. Liu, "Joint Iterative Demodulation and Decoding of Differential Frequency Hopping Signals," *in Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007, pp. 649-652.

16. R. Zhang, K. Liu, "System-by-symbol MAP Detection of Differential Frequency Hopping Signals over Rayleigh Flat Fading Channels," *in Proc. of IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, 2005, pp. 1602-1605.

17. R. Zhang, K. Liu, "Symbol-by-symbol MAP Detection of Differential Frequency hopping Signals for PLC Applications," *in Proc. of IEEE International Symposium on Power-Line Communications and Its Applications (ISPLC)*, 2005, pp. 105-108.

References

[1] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, 2000.

[2] S. Ornager, "View a picture: theoretical image analysis and empirical user studies on indexing and retrieval," *Swedish Library Research*, vol. 2, no. 3, pp. 31–41, 1996.

[3] ——, "Image retrieval: theoretical analysis and empirical user studies on accessing information in images," in *Proceedings of the ASIS Annual Meeting*, vol. 34, 1997, pp. 202–211.

[4] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: ideas, influences, and trends of the new age," *ACM Computing Surveys*, vol. 40, no. 2, pp. 5:1–5:60, April 2008.

[5] L. Guan, Y. Wang, R. Zhang, Y. Tie, A. Bulzacki, and M. Ibrahim, "Multimodal information fusion for selected multimedia applications," *International Journal of Multimedia Intelligence and Security*, vol. 1, no. 1, pp. 5–32, 2010.

[6] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by image and video content: the QBIC system," *IEEE Computer*, vol. 28, no. 9, pp. 23–32, 1995.

[7] C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic, and W. Equitz, "Efficient and effective querying by image content," *Journal of Intelligent Information Systems*, vol. 3, no. 3-4, pp. 231–262, 1994.

[8] J. R. Smith and S.-F. Chang, "Visualseek: a fully automated content-based image query system," in *Proceedings of the ACM international conference on Multimedia*, 1996, pp. 87–98.

[9] A. Pentland, R. W. Picard, and S. Sclaroff, "Photobook: content-based manipulation of image databases," *International Journal of Computer Vision*, vol. 18, no. 3, pp. 233–254, 1996.

[10] J. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R. Jain, and C. Shu, "Virage image search engine: an open frameworkfor image management," in *Proceedings of Storage and Retrieval for Still Image and Video Databases IV*, vol. 2670, San Jose, CA, USA, 1996, pp. 76–87.

[11] Y. Rui, T. Huang, S. Mehrotra, and M. Ortega, "A relevance feedback architecture for content-based multimedia information retrieval systems," in *Proceedings of the Workshop on Content-Based Access of Image and Video Libraries (CBAIVL)*, Washington, DC, USA, 1997, pp. 82–89.

[12] D. Heisterkamp, J. Peng, and H. Dai, "Feature relevance learning with query shifting for content-based image retrieval," in *Proceedings of IEEE International Conference on Pattern Recognition (ICPR)*, 2000, pp. 250–253.

[13] Y. Ishikawa, R. Subramanya, and C. Faloutsos, "Mindreader: query databases through multiple examples," in *Proceedings of the International Conference on Very Large Data Bases (VLDB)*, San Francisco, CA, USA, 1998, pp. 218–227.

[14] S. Aksoy, R. Haralick, F. Cheikh, and M. Gabbouj, "A weighted distance approach to relevance feedback," in *Proceedings of the International Conference on Pattern Recognition (ICPR)*, Barcelona, Spain, 2000, pp. 812–815.

[15] Y. Rui and T. Huang, "Optimizing learning in image retrieval," in *Proceedings of IEEE Conference on Computer Vision Pattern Rrecognition (CVPR)*, vol. 1, 2000, pp. 236–243.

[16] T. Ashwin, N. Jain, and S. Ghosal, "Improving image retrieval performance with negative relevancefeedback," in *Proceeding of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, Salt Lake City, UT ,USA, 2001, pp. 1637 – 1640.

[17] P. Muneesawang and L. Guan, "An interactive approach for CBIR using a network of radial basis functions," *IEEE Trans. Multimedia*, vol. 6, no. 5, pp. 703–716, Oct 2004.

[18] K. H. Yap and K. Wu, "A soft relevance framework in content-based image retrieval systems," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 12, pp. 1557–1568, December 2005.

[19] S. Tong and E. Chang, "Support vector machine active learning for image retrieval," *Proceedings of the ninth ACM international conference on Multimedia*, pp. 107–118, 2001.

[20] Z. Su, H. Zhang, S. Li, and S. Ma, "Relevance feedback in content-based image retrieval: Bayesian framework, feature subspaces, and progressive learning," *IEEE Trans. Image Process.*, vol. 12, no. 8, pp. 924–937, 2003.

[21] G. Giacinto, F. Roli, and G. Fumera, "Comparison and combination of adaptive query shifting and featurerelevance learning for content-based image retrieval," in *Proceedings of the IEEE International Conference on Image Analysis and Processing (ICIP)*, 2001, pp. 422–427.

[22] X. He, O. King, W. Ma, M. Li, and H. Zhang, "Learning a semantic space from user's relevance feedback for image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 1, pp. 39–48, 2003.

[23] S. Uchihashi and T. Kanade, "Content-free image retrieval based on relations exploited from user feedbacks," in *Proc. International Conference on Multimedia and Expo (ICME)*, 2005, pp. 1358 – 1361.

[24] R. Zhang and L. Guan, "A new relevance feedback framework for content-free image retrieval," in *Proceedings of the International Workshop on Multimedia Signal Processing (MMSP)*, 2008, pp. 685–690.

[25] Z. Guo, Z. Zhang, E. P. Xing, and C. Faloutsos, "A max margin framework on image annotation and multimodal image retrieval," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2007, pp. 5–32.

[26] F. Jing, M. Li, H.-J. Zhang, and B. Zhang, "A unified framework for image retrieval using keyword and visual features," *IEEE Trans. Image Process.*, vol. 14, no. 7, pp. 979 – 989, 2005.

[27] X. Anguera, J. Xu, and N. Oliver, "Multimodal photo annotation and retrieval on a mobile phone," in *Proceedings of the ACM International Conference on Multimedia Information Retrieval (MIR)*, New York, NY, USA, 2008, pp. 188–194.

[28] X. Xie and et.al., "Mobile search with multimodal queries," *Proc. IEEE*, vol. 96, pp. 589–601, 2008.

[29] F. Jing, M. Li, H.-J. Zhang, and B. Zhang, "A unified framework for image retrieval using keyword and visual features," *IEEE Trans. Image Process.*, vol. 14, pp. 979–989, 2005.

[30] M. Szummer and R. W. Picard, "Indoor-outdoor image classification," in *Proc. of IEEE International Workshop on Content-Based Access of Image and Video Database*, 1998, pp. 42–51.

[31] A. Vailaya and et.al., "Image classification for content-based indexing," *IEEE Trans. Image Process.*, vol. 10, no. 1, pp. 117–130, 2001.

[32] A. Bosch, A. Zisserman, and X. Muoz, "Scene classification using a hybrid generative/discriminative approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 4, pp. 712–727, Apr. 2008.

[33] F.-F. Li and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2005, pp. 524–531.

[34] L.-J. Li, R. Socher, and F.-F. Li, "Towards total scene understanding: classification, annotation, segmentation in an automatic framework," in *Proceedings of the IEEE*

International Conference on Computer Vision and Pattern Recognition (CVPR), Florida, USA, Jun. 2009, pp. 2036–2043.

[35] F. Monay and D. Gatica-Perez, "On image auto-annotation with latent space models," in *Proceedings of the ACM International Conference on Multimedia*, 2003, pp. 275–278.

[36] ——, "pLSA-based image auto-annotation: constraining the latent space," in *Proceedings of the ACM International Conference on Multimedia*, 2004, pp. 348–351.

[37] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *Proceedings of the ACM International Conference on Research and Development in Informaion Retrieval (SIGIR)*, 2003, pp. 119–126.

[38] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 394–410, 2007.

[39] J. Li and J. Wang, "Read-time computerized annotation of pictures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 6, pp. 985–1002, 2008.

[40] X. J. Wang, L. Zhang, X. Li, and W. Y. Ma, "Annotating images by mining image search results," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1919–1932, 2008.

[41] A. Torralba, "Contextual priming for object detection," *International Journal of Computer Vision*, vol. 53, no. 2, pp. 169–191, 2003.

[42] P. Duygulu, K. Barnard, J. F. G. de Freitas, D. A. Forsyth *et al.*, "Object recognition as machine translation: learning a lexicon for a fixed image vocabulary," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2002, pp. 97–112.

[43] P. Carbonetto, N. Freitas, and K. Barnard, "A statistical model for general contextual object recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2004, pp. 350–362.

[44] D. M. Blei and M. I. Jordan, "Modeling annotated data," in *Proceedings of the ACM International Conference on Research and Development in Informaion Retrieval (SIGIR)*, 2003, pp. 127–134.

[45] K. Barnard, P. Duygulu, D. Forsyth, N. Freitas, D. M. Blei, and M. I. Jordan, "Matching words and pictures," *Journal of Machine Learning Research (JMLR)*, vol. 3, pp. 1107–1135, 2003.

[46] F. R. Kschischang, B. J. Frey, and H. A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 498–519, Feb 2001.

[47] E. Chang, K. Goh, G. Sychay, and G. Wu, "CBSA: content-based soft annotation for multimodal image retrieval using bayes point machines," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 1, Jan. 2003.

[48] J. Li and J. Z. Wang, "Automatic linguistic indexing of pictures by a statistical modeling approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1075–1088, September 2003.

[49] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning object categories from Google's image search," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, vol. 2, 2005, pp. 1816–1823.

[50] B. C. Russell, W. T. Freeman, A. A. Efros, J. Sivic, and A. Zisserman, "Using multiple segmentations to discover objects and their extent in image collections," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006, pp. 1605–1614.

[51] Y. Mori, H. Takahashi, and R. Oka, "Image-to-word transformation based on dividing and vector quantizing images with words," in *Proceedings of International Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999, pp. 405–409.

[52] K. Barnard and D. Forsyth, "Learning the semantics of words and pictures," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, vol. 2, 2001, pp. 408–415.

[53] K. Barnard, P. Duygulu, and D. Forsyth, "Clustering art," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2001, pp. 434–441.

[54] P. F. Brown, V. J. Della Pietra, S. A. Della Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: parameter estimation," *Journal Computational Linguistics - Special issue on using large corpora*, vol. 19, pp. 263–311, Jun 1993.

[55] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, 2000.

[56] K. Mikolajczyk and C. Schmid, "Scale affine invariant interest point detectors," *International Journal of Computer Vision (IJCV)*, vol. 60, pp. 63–86, Oct. 2004.

[57] J. Matas, "Robust wide-baseline stereo from maximally stable extremal regions," *Journal of Image and Vision Computing*, vol. 22, no. 10, Oct 2004.

[58] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision (IJCV)*, vol. 60, pp. 91–110, Nov. 2004.

[59] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering object categories in image collections," Massachusetts Institute of Technology Computer Science and Artificial Intelligence Laboratory, Tech. Rep. MIT-CSAIL-TR-2005-012, Feb. 2005.

[60] V. Lavrenko, R. Manmatha, and J. Jeon, "A model for learning the semantics of pictures," in *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)*, 2003.

[61] S. L. Feng, R. Manmatha, and V. Lavrenko, "Multiple bernoulli relevance models for image and video annotation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2004, pp. 1002–1009.

[62] V. Lavrenko and W. B. Croft, "Relevance based language models," in *Proceedings of the ACM International Conference on Research and Development in Information retrieval (SIGIR)*, 2001, pp. 120–127.

[63] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, 1998, pp. 275–281.

[64] O. Maron and T. L. Pérez, "A framework for multiple-instance learning," in *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)*, Cambridge, MA, USA, 1997, pp. 570–576.

[65] A. Kalai and A. Blum, "A note on learning from multiple-instance examples," *Machine Learning*, vol. 30, no. 1, pp. 23 – 30, January 1998.

[66] I. Biederman, R. J. Mezzanotte, and J. C. Rabinowitz, "Scene perception: detecting and judging objects undergoing relational violations," *Cognitive psychology*, vol. 9, pp. 411–429, 1983.

[67] I. Biederman, R. C. Teitelbaum, and R. J. Mezzanotte, "Scene perception: A faliure to find a benefit of expectancy or famliiarity," *Journal of Experimental Psychology: Human Learning And Memory*, vol. 14, no. 2, pp. 143–177, April 1982.

[68] M. Bar., "Visual objects in context," *Nature Reviews Neuroscience*, vol. 5, no. 8, pp. 617–629, 2004.

[69] M. C. Potter, "Meaning in visual search." *Science*, vol. 187, no. 4180, pp. 965–966, Mar. 1975.

[70] D. Navon, "Forest before trees: the precedence of global features in visual perception," *Cognitive Psychology*, vol. 9, no. 3, pp. 353–383, Jul. 1977.

[71] R. A. Rensink, J. K. O'Regan, and J. J. Clark, "To see or not to see: The need for attention to perceive changes in scenes," *Psychological Science*, vol. 8, no. 5, pp. 368–373, Sep. 1997.

[72] A. Torralba, K. Murphy, and W. Freeman, "Using the forest to see the trees: object recognition in context," *Communications of the ACM, Research Highlights*, 2009.

[73] L. Wolf and S. Bileschi, "A critical view of context," *International Journal of Computer Vision*, vol. 69, no. 2, pp. 251–261, August 2006.

[74] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, "Objects in context," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2007.

[75] J. Yuan, J. Li, and B. Zhang, "Exploiting spatial context constraints for automatic image region annotation," in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2007, pp. 595–604.

[76] A. Singhal, J. Luo, and W. Zhu, "Probabilistic spatial context models for scene content understanding," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2003, pp. 235–241.

[77] A. Oliva and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, pp. 145–175, 2001.

[78] B. Clarkson and A. Pentland, "Framing through peripheral perception," in *Proceedings of IEEE Conference on Image Processing (ICIP)*, 2000, pp. 38–41.

[79] A. Torralba, "Contextual modulation of target saliency," in *In Advances in Neural Information Processing Systems (NIPS)*, vol. 14, 2002, pp. 1303–1310.

[80] A. Torralba, K. Murphy, and W. Freeman, "Using the forest to see the trees: exploiting context for visual object detection and localization," *Communications of the ACM, Research Highlights*, vol. 53, no. 3, pp. 107–114, 2010.

[81] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-6, no. 6, pp. 721–741, Nov. 1984.

[82] S. Z. Li, *Markov random field modeling in image analysis.* Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2001.

[83] K. P. Murphy, Y. Weiss, and M. I. Jordan, "Loopy belief propagation for approximate inference: An empirical study," in *In Proceedings of Uncertainty in AI*, 1999, pp. 467–475.

[84] J. Besag, "Statistical analysis of dirty pictures," *Journal of Applied Statistics*, vol. 20, no. 5, pp. 63–87, 1993.

[85] A. Berger, "The improved iterative scaling algorithm: a gentle introduction," 1997.

[86] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2001, pp. 282–289.

[87] X. H., R. S. Zemel, and Miguel, "Multiscale conditional random fields for image labeling," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2004, pp. 695–702.

[88] S. Kumar and M. Hebert, "Discriminative random fields: A discriminative framework for contextual interaction in classification," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, vol. 2, 2003, pp. 1150–1157.

[89] A. Quattoni, M. Collins, and T. Darrell, "Conditional random fields for object recognition," in *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)*, vol. 17, 2004, pp. 1097–1104.

[90] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1988.

[91] C. Zitnick, "Computing conditional probabilities in large domains by maximizing rényi's quadratic entropy," Ph.D. dissertation, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, May 2003.

[92] R. Zhang, K. Wu, K.-H. Yap, and L. Guan, "A collaborative Bayesian image annotation framework," in *Proceedings of the Pacific-Rim Conference on Multimedia (PCM)*, 2008, pp. 348–357.

[93] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics*, vol. 23, pp. 309–314, 2004.

[94] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, 2002.

[95] M. A. Stricker and M. Orengo, "Similarity of color images," in *Proceedings of the SPIE Conference on Storage and Retrieval for Image and Video Databases III*, vol. 2420, March 1995, pp. 381–392.

[96] J. R. Smith and S.-F. Chang, "Automated binary texture feature sets for image retrieval," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, 1996, pp. 2239–2242.

[97] "MSRC v2.0: http://research.microsoft.com/en-us/projects/objectclassrecognition/."

[98] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: a large dataset for non-parametric object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1958–1970, 2008.

[99] C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[100] B. S. Manjunath, J. R. Ohm, V. V. Vinod, and A. Yamada, "Color and texture descriptors," *IEEE Trans. Circuits and Systems for Video Technology, Special Issue on MPEG-7*, vol. 11, no. 6, pp. 703–715, Jun 2001.

[101] B. S. Manjunath and W. Ma, "Texture features for browsing and retrieval of image data," *IEEE Trans. Pattern Anal. Machine Intell. (PAMI - Special issue on Digital Libraries)*, vol. 18, no. 8, pp. 837–42, Aug 1996.

[102] G. Lu, "Indexing and retrieval of audio: A survey," *Springer Netherlands*, vol. 3, pp. 269–290, October 2004.

[103] P. Quelhas and J. marc Odobez, "Natural scene image modeling using color and texture visterms," in *Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR)*, Tempe, AZ, USA, Jul. 2006, pp. 411–421.

[104] R. Lienhart, S. Romberg, and E. Hörster, "Multilayer pLSA for multimodal image retrieval," in *Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR)*, 2009, pp. 9:1–9:8.

[105] K. E. van de Sande, T. Gevers, and C. G. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1582–1596, Sep. 2010.

[106] T. Hofmann and J. Puzicha, "Statistical models for co-occurrence data," Massachusetts Institute of Technology Artificial Intelligence Laboratory, Tech. Rep. AIM-1625, Feb. 1998.

[107] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2009 (VOC2009) Results," http://www.pascal-network.org/challenges/VOC/voc2009/workshop/index.html.

[108] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: a database and web-based tool for image annotation," *International Journal of Computer Vision*, vol. 77, no. 1-3, May 2008.