EXPLOITING MULTIMEDIA CONTENT: A MACHINE LEARNING BASED APPROACH

EHTESHAM HASSAN

DEPARTMENT OF ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY DELHI

NOVEMBER 2012

EXPLOITING MULTIMEDIA CONTENT: A MACHINE LEARNING BASED APPROACH

by

EHTESHAM HASSAN

Department of Electrical Engineering

Submitted

in fulfillment of the requirements of the degree of

Doctor of Philosophy

to the

INDIAN INSTITUTE OF TECHNOLOGY DELHI

NOVEMBER 2012

Certificate

This is to certify that the thesis titled "**EXPLOITING MULTIMEDIA CONTENT: A MACHINE LEARNING BASED APPROACH**" being submitted by **EHTESHAM HASSAN** to the Department of Electrical Engineering, Indian Institute of Technology Delhi, for the award of the degree of Doctor of Philosophy, is a record of bona-fide research work carried out by him under our guidance and supervision. In our opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree. The results contained in this thesis have not been submitted to any other university or institute for the award of any degree or diploma.

Prof. Madan Gopal

Department of Electrical Engineering

India Institute of Technology Delhi

New Delhi - 110016

Prof. Santanu Chaudhury

Department of Electrical Engineering

India Institute of Technology Delhi

New Delhi - 110016

To my family

Acknowledgements

First of all my ethereal indebtedness and recognition goes to the Almighty that He steers me in my whole walk of life.

The successful completion of this thesis is due to the mode of supervision, timely encouragement and efficient guidance received from my supervisors Professor Madan Gopal, and Professor Santanu Chaudhury. My deep appreciation and bounteous praise goes to my supervisors for their patience and assistance in settling my fuzzy ideas, confusions and short comings. The thesis would not achieved its present form without their continuous erudite help and conscientious support. My heartfelt thanks and regards to Santanu sir for all his financial support and freedom which never let me bother for any publication charge and other requirements. I acknowledge with gratitude the benevolence of my seniors, friends and the staff of Control lab and Multimedia lab who have been ever helpful and supportive during the course of this thesis. Last but not the least my benevolent gratitude, credit affirmation goes to my family members for their plunk for, plump for and scrupulous sustainability and unsparing and unstinted counselling, steering and guidance throughout the course of this research project and in whole walk of life.

Abstract

This thesis explores use of machine learning for multimedia content management involving single/multiple features, modalities and concepts. We introduce shape based feature for binary patterns and apply it for recognition and retrieval application in single and multiple feature based architecture. The multiple feature based recognition and retrieval frameworks are based on the theory of multiple kernel learning (MKL). A binary pattern recognition framework is presented by combining the binary MKL classifiers using a decision directed acyclic graph. The evaluation is shown for Indian script character recognition, and MPEG7 shape symbol recognition. A word image based document indexing framework is presented using the distance based hashing (DBH) defined on learned pivot centres. We use a new multi-kernel learning scheme using a Genetic Algorithm for developing a kernel DBH based document image retrieval system. The experimental evaluation is presented on document collections of Devanagari, Bengali and English scripts.

Next, methods for document retrieval using multi-modal information fusion are presented. Text/Graphics segmentation framework is presented for documents having a complex layout. We present a novel multi-modal document retrieval framework using the

segmented regions. The approach is evaluated on English magazine pages. A document script identification framework is presented using decision level aggregation of page, paragraph and word level prediction. Latent Dirichlet Allocation based topic modelling with modified edit distance is introduced for the retrieval of documents having recognition inaccuracies. A multi-modal indexing framework for such documents is presented by a learning based combination of text and image based properties. Experimental results are shown on Devanagari script documents.

Finally, we have investigated concept based approaches for multimedia analysis. A multi-modal document retrieval framework is presented by combining the generative and discriminative modelling for exploiting the cross-modal correlation between modalities. The combination is also explored for semantic concept recognition using multi-modal components of the same document, and different documents over a collection. An experimental evaluation of the framework is shown for semantic event detection in sport videos, and semantic labelling of components of multi-modal document images.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The exponential growth in multimedia contents and novel innovations in accessibility options have created the demand for sophisticated analysis tools for efficient content management. In particular, there exists a need for developing more effective methods for indexing, searching, categorizing and organizing this information. With the explosion in database sizes, retrieval has been a key challenge in the multimedia database management. The challenge in retrieving desired multimedia information is multi-dimensional. The process of feature extraction for data representation needs to be selective and invariant to sensor errors. The similarity matching between data instances significantly affects the success in organizing data or finding relevant results. The matching needs to address the non-linear relationship between attributes across instances as well as computational constraints. Retrieval over large databases requires an efficient indexing scheme which needs to be accurate within error bounds even if efficiencies requires the use of approximations. In addition, the scheme should be scalable and adaptive to address the rapidly

increasing volume of data. In practice, keyword based querying is the preferred user query mode where query words express the user's intent. The non-linear relationship between user's intent i.e. high level concepts and low-level input features results in the semantic gap problem. A typical content management system encounters multimedia data containing multi-modal information e.g. videos having audio, video and ticker text, images in running text and image collections having textual description attached as tags. Multiple modalities contribute unique contextual information, the fusion of which can improve the understanding of semantics of multimedia data, helping in bridging the 'semantic gap'.

Various machine learning methods have been developed to provide the technology for different applications. The unique capability of abstracting the semantics of data over large collections make machine learning methods ideal for multimedia content exploitation. Through learning, structural insight of the content space can be extracted for representation in a comprehensive form. The representation helps in tasks like recognition, annotation, retrieval and compression. The learning algorithms exploit distinct attributes of the content applicable for different tasks.

1.1 Scope and Objective

In general, multimedia content belongs to two classes: temporal media and non-temporal media. The temporal media also referred as to dynamic media has associated time information in which content changes with the passing of time. Example of such media types include video, speech, audio and animation. The representation of the content of non-

temporal media or static media does not change with respect to time e.g. text, images and graphics. A document has this information in independent form: pure text, image or audio file, or different types of information in coexistence, i.e., documents having multi-modal information. Example of such documents include text appearing on an image canvas, text appearing in a video stream as closed captions and ticker text, video file having audio information. Many documents also have external but associated multi-modal information, e.g., text annotations and tags, available with video, and audio files. In this thesis, we focus on developing machine learning based techniques for dealing with different multimedia components, in individual form as well as in co-occurring combinations, for retrieval applications.

Efficient indexing simplifies the management and preservation of a large collection of multimedia documents by assigning a unique index to documents having similar content, or expressing similar semantics. The existing multimedia document indexing methods generate the document indices using the extracted features characterizing the underlying content. Typical feature extraction methods exploit the low-level content attribute such as intensity, color and shape. Uniqueness of the information provided by different attributes contribute varying invariance and robustness to different feature sets. Feature based representation being the primary information channel for understanding the semantics of content, define the performance quality of multimedia retrieval systems. Also, the multiple modalities existing in documents provide complementary information for improved understanding of the semantics of underlying content. In this thesis, we concentrate on the application of information from multiple features and multiple modalities for the recogni-

tion and retrieval of multimedia documents. Learning based frameworks are defined for multi-modal retrieval of documents by integrating the feature space representation, as well as concept based integration for semantic linking and categorization.

The text in multimedia documents appears in two forms: ASCII text documents and, imaged forms of text information, i.e., document images (Refer figure 1.1). Text information in multimedia documents also appear with embedded graphics and image information such as a ASCII text document having image content, or a document image having embedded graphics or an image component. As one category of data in the thesis, we have considered documents having text in imaged form with embedded graphics/image information. Example of such documents include old books, historical documents, and manuscripts. Large amount of such documents is available in various libraries across the world. We explore different challenges and problems associated with management of such documents and present learning based methods for managing such collections. The preservation and management of document image collection follows two procedures. First procedure converts the document image in digitized form, and applies existing text based indexing and retrieval methods for accessing the collection. The approach poses following major challenges:

- Robust recognition technology for document conversion to the digitized form.
- Efficient document retrieval method for accessing the collection.

The second approach uses the image based properties of textual and image content for indexing the collection. The indexing framework is required to accurately group the

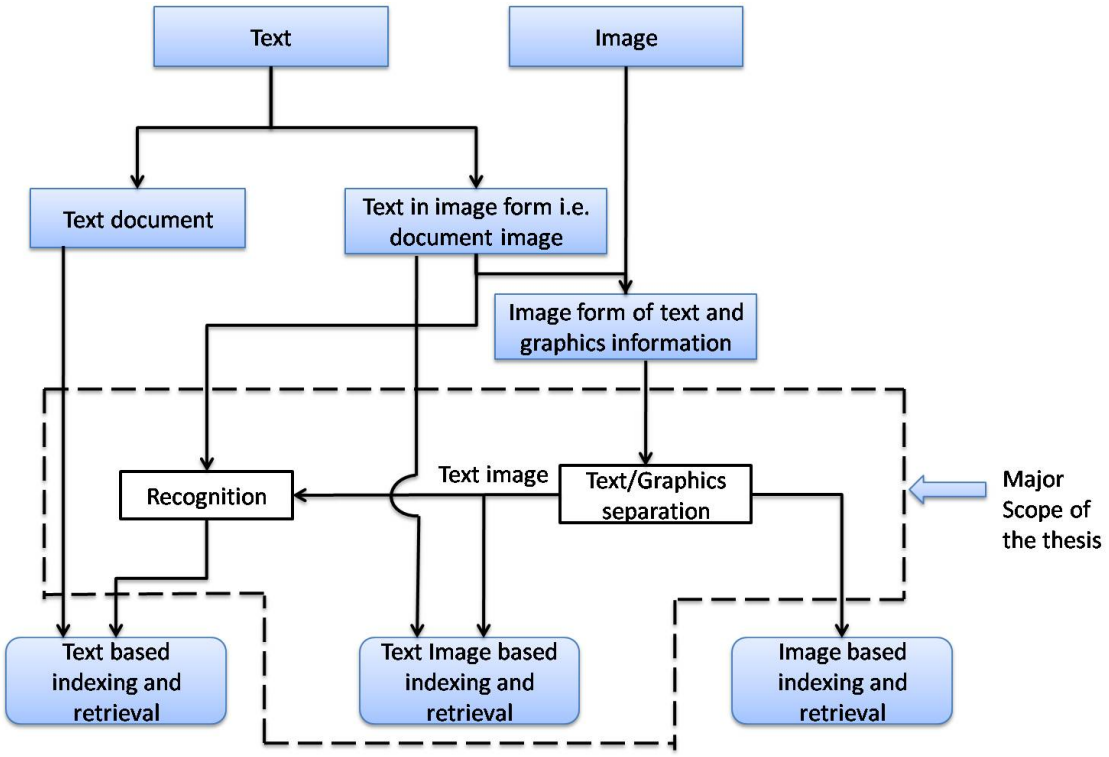


Figure 1.1: Text based multi-media content analysis paradigm

documents having similar content in the generated indexing space. In the case of document images having multi-modal information, the separation of text and image/graphics content is a prior requirement for both procedures. Subsequently, the documents are indexed using text or image based methods. The following challenges associated with the management of document images have been considered in this thesis.

- Image based indexing requires the document content representation in a precise and concise form. An efficient matching algorithm is required for measuring the similarity of the documents. Similarity based matching in a large collection of documents is computationally costly. The method should address the practical requirement of

fast retrieval with sub-linear complexity.

- Text document indexing and retrieval has been well researched and studied by information retrieval community. However, the application of text based retrieval techniques in OCR'ed document collection does not always guarantee satisfactory performance. In this direction, the indexing methods is required to be invariant to recognition inaccuracies.
- Document images having multi-modal information require segmentation of text and graphics regions. Here, a robust classifier is required to correctly identify different type of image regions by the exploitation of content and context information available in the document. The text information also appears in multiple scripts. In this case, the document processing needs the text region to be classified based on script using the image properties.

The multimedia documents having image information have been preferred area of computer vision researchers. The development in multimedia applications have generated large amount of documents having multi-modal content e.g. images with textual description, videos with textual description and audio. The scenario has created the need for new methods as the multi-modal information could significantly improve the semantic understanding of data for different applications. Specifically, we address following problems in this thesis

- Feature representation for binary patterns.
- Indexing and retrieval of documents using single and multiple feature.

- Multi-modal information integration and retrieval of document images.

The user access to multimedia database requires the retrieval to focus on document representation. The content based retrieval describes the information need by query example. Subsequently, the example is compared with all the database documents for retrieving the relevant results. The text based retrieval present more convenient approach for query expression. The approach extends the applicability of well established text document retrieval methods for multimedia retrieval. In this direction, the recent developments in multimedia retrieval have used concept based retrieval. Concept based retrieval represents the documents in terms of automatically detectable semantic concepts. Here, the retrieval does not just search the appearance of the query concept, but also exploits the context of the query for correlating with the multimedia documents. Concepts may represent visibly apparent semantic attribute of documents, or hidden semantics, e.g., latent context underlying the document space which makes their appearance modality independent. The modality independence, therefore, simplifies the approach for multi-modal document retrieval. As the concluding part of this thesis, we investigate the problem of concept learning, and cross-modal modelling for multimedia document management.

1.2 Major Contributions of the Thesis

The work presented in this thesis propose a set of adaptive techniques for processing multimedia content to meet semantic need of users. A set of feature extraction techniques have been proposed which can efficiently capture properties of multimedia content. An

indexing scheme for multimedia content is developed which has the unique capability to learn appropriate feature combination for meeting the objective directed retrieval targets. New schemes for multi-modal feature combination for concept based interpretation, classification and retrieval has been designed. To summarize, the major contributions of the present work are as follows:

1. Multiple feature based recognition framework for binary patterns such as characters, primitives and symbols: Learning based feature combination for OCR and symbol recognition.
   (a) The framework presents an application of multiple kernel learning in a decision directed acyclic graph architecture for large category classification problems.
   (b) A novel shape descriptor for binary pattern feature representation is proposed. The framework is used for recognizing the characters and primitives of Indian scripts using the shape descriptor and other feature representations. The framework is also evaluated for recognizing the symbols from MPEG7 shape dataset.
2. The learning framework for multimedia document indexing and retrieval using single and multiple feature definitions: The application of distance based hashing for document indexing using single, and multiple features using a multiple kernel learning formulation for retrieval.
   (a) The word based document indexing and retrieval framework is presented by using the distance based hashing function for generating the indexing model.

The distance based hashing functions are defined using the learned pivot points. The applicability of the framework is shown for following feature representations and similarity matching methods: 1). Shape descriptor with Euclidean distance based similarity, 2). String like word image representation using edit distance based similarity.

(b) A novel multi-probe hashing framework for binary mapping functions is proposed. The applicability of the framework is shown for distance based hashing functions.

(c) A novel multiple kernel learning formulation for retrieval problems is presented by using the kernel distance based hashing. The genetic algorithm based optimization framework learns the optimal kernel for indexing as parametrized linear combination of base kernels. The framework, therefore, provides a principled approach for using multiple features retrieval problem. The framework is evaluated for word based document indexing and retrieval using multiple feature representations.

3. Methods for multi-modal information integration and application for document retrieval: Identification for multi-modal regions from document images, script identification, multi-modal retrieval, and composite indexing of multi-modal documents.

(a) A novel document image segmentation scheme for a document having overlapped text and graphics information is designed. A multi-modal document image retrieval framework is proposed using the information for text and graph-

ics regions. The framework applies a multiple kernel learning formulation for retrieval for generating the composite document index using different modalities.

(b) A fast and robust script identification framework is proposed for bi-lingual documents having random usage of scripts at page, paragraph, and word level.

(c) The Latent Dirichlet allocation based retrieval framework for documents having recognition errors is proposed. The framework presents latent topic learning adaptive to recognition errors for robust semantic indexing of OCR recognized documents. For improved retrieval of such documents, a word based multi-modal document indexing framework is proposed by combining OCR recognized text and image based representation using learning.

4. Frameworks for multimedia content analysis framework using uni/multi modal concepts for recognition and retrieval applications: Semantic concept learning using multiple features, and integration by probabilistic modelling based multi-modal analysis.

(a) A concept learning framework has been developed for multiple feature based low-level content representation. The framework uses multiple kernel learning for recognition. The efficacy of the framework is demonstrated for posture based conceptual annotation of Indian classical dance images. We also evaluate the framework for learning the LSCOM concepts. A concept based image retrieval framework is proposed using multiple kernel learning.

(b) Multi-modal document retrieval framework by learning the conceptual linkage across documents of multiple modalities, or multi-modal components of same document. The framework proposes conditioned topic learning by combining generative and discriminative modelling. The combinatorial form of generative and discriminative modelling is subsequently applied for the recognition of multi-modal concepts for semantic categorization of documents. The framework is described for an event detection application in sports videos. Also, the application of framework is shown for semantic categorization of multi-modal documents.

1.3 Layout of the thesis

The thesis is organized in eight chapters. Chapter 2 presents the state-of-the-art of existing contributions in multimedia content analysis using learning based approach as specific to the contributions described in section 1.2.

Chapter 3 presents classification framework for binary patterns by applying learning based combination of multiple features. The class of binary patterns i.e. character/primitive and symbol images are considered for evaluation. Novel shape based feature representation for binary patterns is introduced. The classification framework presents decision directed acyclic graph (DDAG) based arrangement of binary multiple kernel learning classifiers for large class problem. The experimental evaluation is performed on Gujarati and Bengali character/primitive collection, and MPEG-7 symbol dataset.

Chapter 4 presents word based indexing framework for document image collections. First, a shape based feature representation for word images is introduced. The representation is used for developing a document indexing framework using distance based hashing. The multi-probing framework using distance based hashing is presented for reducing the size of the indexing data structure. Subsequently, a string like word representation is introduced for developing a document indexing framework using edit distance based hashing. The experimental evaluation of presented concepts is shown on Indian and Latin script document image collections.

Chapter 5 introduces the kernel distance based hashing. The kernel distance based hashing is applied for defining a multiple kernel learning formulation for the retrieval problem. The initial evaluation of the concept is shown for a handwritten digit dataset. Subsequently, the multiple kernel learning formulation is applied for developing a word based document indexing framework using multiple feature representations. The experimental evaluation of retrieval framework is presented on Indian and Latin script document image collections.

Chapter 6 presents a conceptual framework for segmenting multi-modal document images into text, graphics and background regions. The framework is subsequently applied for multi-modal retrieval application for document images. Subsequently, a script recognition framework for mixed script document images is presented. The final section of the chapter presents retrieval framework for text documents having recognition errors. First, a topic modelling based document indexing framework invariant to recognition errors is presented. Subsequently, a multiple kernel learning based framework for improved

retrieval of erroneous text documents is presented.

Chapter 7 introduces a concept learning framework using multiple kernel learning based feature combination. The applicability of the framework is shown for semantic concept based image annotation and retrieval. Next, a conditioned topic modelling for learning the association between multi-modal semantic concepts is presented. The evaluation is presented for multi-modal document retrieval having image and text information. The final section presents new formulation of probabilistic techniques used for conditioned topic learning for multi-modal concept recognition. The evaluation is shown for sport event detection and semantic concept detection in multi-modal documents.

Chapter 8 presents the conclusions and scope for further work in the area.

Chapter 2

Learning for Multimedia Content Management: Revisited

The advances in machine learning, pattern recognition, information retrieval and computer vision have contributed significantly towards learning based analysis and management of multimedia content. However, the growing rate of content generation and development of novel applications have rendered the multimedia content management problems challenging. In the following discussion, we present a brief survey of existing contributions towards multimedia analysis and management. We primarily concentrate on the problem of multimedia document indexing and retrieval, and analyse the state-of-the-art with the machine learning perspective. The multimedia applications extract the information content from the documents by the process of feature extraction. The feature extraction uses low-level content attributes e.g. color, shape and texture for defining numeric representation of the documents. However, the robustness and invariance properties of these features

for different classes of examples are always questionable. The effectiveness of feature based content representation may be improved by exploiting the complementary information represented by different feature extraction algorithms. Further, the semantic gap between conceptual requirement of users and content level representation of multimedia content requires extended usage of machine learning technique for multimedia data management. We begin the review with a discussion of the application of multiple features for multimedia analysis. Subsequently, the learning based methods for multimedia retrieval are discussed. Finally, retrieval and analysis methods specific to different modalities are briefed.

2.1 Analysis of Multiple Feature based Recognition and Retrieval

The effectiveness of a multimedia document recognition and retrieval system is dependent on the discriminating capability of the feature set used. The extracted feature set from the document must exhibit a high level of invariance for similar examples, with a high level of discriminative power across different examples. Nevertheless, a single feature description can not guarantee uniform discriminative capability across the dataset in feature space as invariance properties of feature representations vary across the example categories. Therefore, an adaptive combination of a set of diverse and complementary features based on different object attributes e.g. color, shape and texture improves the discriminative power of resultant feature set. We have seen that primarily two strategies have been used

for combining different features for improved recognition. The first strategy referred as early fusion addresses the combination in feature space itself, i.e., feature level fusion [7, 228, 184, 159]. The second strategy referred as late fusion follows classifier level fusion of features, i.e., the final prediction score is generated by combining prediction scores learned over individual features [154, 350, 108, 74]. In this context, [265] and [162] presented theoretical and empirical evaluations of different classifier fusion methods. Recent work has also explored learning based feature selection approach for multiple features based recognition [231, 310, 54]. At this point, we review the concept of Multiple kernel learning (MKL) for classification. The Support vector machine (SVM) is widely accepted state-of-the-art classifier based on kernel learning. SVM is a binary classifier which defines quadratic optimization problem to learn the maximum margin hyperplanes separating the classes in high dimensional kernel space. Recent research in SVMs and other kernel learning methods have shown that using multiple kernels instead of a single kernel can improve the interpretation of decision function as well as classifier performance. This problem is solved in the MKL framework, where the optimal kernel for classification is learned by combining a set of base kernels through the process of learning. Additionally, given that a typical learning problem often involves multiple, heterogeneous data source, the MKL provides a principled approach for combining information from such data sources.

Existing MKL Formulations and Application for Feature Combinations

The most natural approach for MKL is to consider linear combinations of kernels. The combination of kernels is defined as $K_c = \sum_{k=1}^{M} \eta_k K_k$, where $\boldsymbol{\eta}$ are kernel weight param-

eters. In recent research, many MKL formulations have been proposed [252, 293, 168]. In this context, considering an unweighed base kernel set, i.e., direct addition of kernels is the simplest approach. In [172], Lazebnik *et al.* presented an extensive discussion on different combinations of features for texture analysis. However, unweighed addition gives equal preference to all kernels which may not be optimal. The weighted combination of kernels present a more logical approach. The weights in this case define the importance of the kernels for discrimination. The MKL algorithm learns these weight parameters from the training data. Lanckriet *et al.* [168] have proposed a conic combination of kernel matrices by formulating a quadratically constrained quadratic program. In [293], Sonnenburg *et al.* formulated the MKL problem as a semi-infinite linear program where the SVM parameters and kernel weights are learned simultaneously in a two-step process. First, the SVM parameters are learned by a standard solver following the optimization of kernel weights by solving a cutting plane algorithm. In [226], the authors have considered a weighted combination of the color, shape and texture features for flower recognition. The weights are learned by optimizing the performance measure. In [111], Gehler and Sebastian presented an extensive evaluation of different kernel combination rules in MKL for object recognition. Campos *et al.* [65] have combined six different features using MKL for character recognition in natural images. Song *et al.* [292] have applied MKL for locality specific feature combination for action recognition in videos. In [309], feature selection from satellite images is performed by learning a linear combination of kernels representing contextual and textural features. The recent work by Lan *et al.* [166], applied MKL based feature combination for event detection.

2.1.1 Multiple Feature based Retrieval

The use of multiple features for various indexing and retrieval applications is common. The concatenation of different features for document representation is the simplest approach. In this direction, one of the earliest work presented in [56] used concatenated texture features for image retrieval. In [55], a combination of color and texture features for image retrieval is done by concatenating the normalised features. The normalization is performed by feature dimension and standard deviation to reduce the effect of different feature dimensions and variances. In [117], image representation is defined by concatenating the bag-of-words histogram computed for different features. Bai *et al.* [20] concatenated a set of topological shape features for word shape coding for word based document retrieval. The concatenation increases the resulting feature dimension, thereby increasing the retrieval complexity. Additionally, the resulting feature set does not guarantee improved discriminative capability because the complementary information of different features is not optimally exploited. The linear combination of descriptors is another simple approach for feature combination. The preferred approach in this direction has been to create separate index structures for different descriptors, and combine them at the time of query retrieval. In [77], color and texture features have been combined for image retrieval by adding the query distance to example images using both the features. The authors in [106] have shown improvements in retrieval results by considering the geometric mean of similarity measures computed with different features. In [247], a heuristic based feature combination approach is defined for image retrieval which combines the similarity scores obtained for color, texture and edge

histogram features by a non-linear function. Rath and Manmatha [255] applied a linear combination approach for word based document retrieval by merging the feature sets with uniform weights. However, the learning based methods for feature combination provide a more constructive approach than heuristic based methods. Learning based feature combination has shown improved results for various classification and recognition problems. Dasigi *et al.* [63] presented one of the earliest work on learning based feature combination for text classification. The neural network based supervised learning framework is defined for combining the low dimensional document representation obtained by Latent Semantic Analysis. Lin and Bhanu [182, 183] have explored the application of Genetic Programming for learning the combination of different features for object recognition problem. In [277], neural network learning is applied to perform non-linear dimensionality reduction over a combination of pitch, timbre and rhythm features for music retrieval. In [23], feature combination at local neighbourhood level is performed for face image retrieval. The feature combination is learned as Genetic Algorithm based learning for local regions and user feedback based learning for region weights for confidence measurement.

2.2 Multimedia Database Indexing

Unlike conventional databases, multimedia databases preserve the documents as collection of features characterizing the distinct attributes of inherent content. The user information requirement is provided by correlating the query with database documents using feature based similarity. The development of multimedia database application consists of three

primary problems: Feature extraction, Similarity matching and a Retrieval scheme. Significant amounts of research effort have been contributed towards addressing these problems [33, 175, 78, 288].

The research on feature extraction methods have contributed several local descriptors which has shown improved robustness to noise, partial visibility and occlusion problem [25, 62, 192, 169, 332]. The local descriptors based representation primarily uses a bag-of-words model which significantly reduces the feature dimension and helps in indexing a large-scale image/video collection. Primarily, multimedia retrieval problem has been addressed using two approaches. Content based retrieval, where user provides an example object which is matched to the object collection. The similarity establishment between two images computes matching either at pixel, feature and object level. In [288, 78], several methods computing similarity at different levels e.g. similarity between features, object silhouettes, structural features, and similarity at the semantic level has been discussed. The recent effort in similarity matching has proposed distance metric learning for image matching that preserves the distance relation among the training data [347, 136, 105, 132]. The second approach for retrieval accepts the query in text form representing the semantics of user information need. The retrieval is subsequently performed by matching query to the semantics of indexed documents [4]. The process uses set of intermediate semantic concepts to describe frequent visual content of the multimedia document. The process therefore attempts to bridge the semantic gap between high-level image semantics and low-level features describing the visual property of the content. In [189], a detailed description on existing methods for semantics based image retrieval is presented. In [291] presented

detailed description of semantic concept based video retrieval. Liu *et al.* [188] proposed a decision tree for learning the semantic concepts related to image. Recent work by Jiang and Ngo [149] proposed ontology based semantic video indexing by combining a soft-weighing of visual-words with respect to its linguistic meaning and constraint-based earth mover's distance for measuring the linguistic similarity of visual words. In the following section, we discuss the existing retrieval methods primarily concentrating on hashing based indexing schemes.

2.3 Hashing based Indexing Schemes

Traditionally, nearest neighbour search has been the most preferred retrieval algorithm because of its simplicity and robustness [28]. However the linear time search complexity $(O(n))$ is practically unacceptable in modern retrieval applications handling large amount of high dimensional multimedia data. The approximate nearest neighbour search methods give efficient solution to this problem. These algorithms achieve sub-linear search complexity with trade off in terms of marginal decrease in accuracy [141]. Hashing based indexing for various applications is a popular research problem. The scheme generates object indices by projection on a lower dimensional hash space. The process generates a hash table containing multiple buckets. Each bucket represents group of objects corresponding to an unique index. The retrieval process includes a query index generation by applying the same mapping function to query. The relevant documents are obtained by performing similarity search over the objects in bucket corresponding to query index.

The work presented in [209] demonstrated one of the earliest application of Geometric hashing for handwritten document indexing. Locality sensitive hashing (LSH) introduced by Indyk and Motwani is state-of-the-art method for finding similar objects in large data collection [8]. LSH solves the approximate nearest neighbour search in $O(n^{1/1+\epsilon})$ for $\epsilon \geq 0$, by projecting the high dimensional data points to a low dimensional hash space with high probability that similar points are mapped to same location. In recent years, many applications have applied LSH for performing similarity search in high dimensional spaces [126, 276, 327, 206]. The LSH based indexing assumes the uniform distribution of objects in feature space for hashing. In this case, the retrieval success probability is increased by generating multiple hash tables and collecting objects corresponding to query buckets of all hash tables for similarity search. However, the generation of multiple hash tables increases space complexity of data structure containing the indexing information. Additionally, the assumption for uniform distribution is not satisfied for most of the practical applications irrespective of theoretical bound on search complexity. In this direction, recent work by Muja and Lowe [221] have experimentally shown the superior performance of clustering based trees for retrieval in comparison with LSH. In recent effort towards reducing the size of LSH data structure, the concept of multi probe hashing has been proposed [194]. Multi-probing reduces the required number of hash tables by increasing the utilization of single hash table. Multi-probe hashing selects more than one bucket from each hash table for probing, therefore increasing the usability of hash tables. In this way, multi-probing considerably reduces the number of hash tables required for high success probability. However the definition of locality sensitive hashing function requires

information about the object space. The Distance based hashing presents another hashing based indexing scheme by preserving the inter object distances in hash space [311]. The objective is object grouping in the hash space based on their similarity measured in terms of defined distance metric. The applicability of DBH has not yet been widely explored for different practical applications. Also, multi-probing framework can not be directly extended for DBH because of the different characteristic of mapping functions. Additionally, recent developments in hashing have presented learning based formulations for defining the hashing function [97, 321, 343, 249, 320]. Kernel matrix in learning methods represents symmetric, positive semidefinite matrix that encodes relative position of all the data points in the feature space. It has been extensively used as the preferred similarity measure for various recognition problems. In this direction, Kulis and Grauman [158] have extended the LSH to kernel space to accommodate the given kernel matrix of training data instead of multidimensional vector representations. Mu et al. [220] have defined quadratic programming based formulation for learning the hashing function with precomputed kernel space. Liu et al. [186] proposed kernel based supervised hashing formulation based on code inner products which employs greedy gradient descent algorithm for solving the hash functions. More recently, the novel concept of hypersphere based hashing is introduced in [130]. The authors have defined spherical Hamming distance for hypersphere- based binary coding, and designed an iterative optimization process for learning independent hashing functions for balanced partitioning of object space.

2.4 Learning for Indexing and Ranking

Ranking in machine learning addresses the class of retrieval problem intend to order the documents based on relevance to the query. Example of such applications include document retrieval and summarisation, collaborative filtering, meta-search and machine translation [185, 176]. From the learning viewpoint, there are two primary challenges: ranking, and fusion. The first simply refers to ranking, while the second refers to aggregation of ranked results from different sources. Nearest neighbour search presents a simple approach for solving ranking problem. In general pairwise preferences have been preferred for defining ranking systems [102, 131, 37, 342]. Such an approach selects pairs of documents from ranked training data as training instances. Each pair is assigned a label based on the relevance of one document with respect to the other. This is used to optimize the ranking problem. In particular for multimedia documents, [94] presented a method by using the MPEG-7 description for retrieval and ranking. In [272], Shao *et al.* discussed ranking in vintage tree based image indexing. He *et al.* [127] proposed manifold learning for ranking by evaluating the relevance of query with example images by exploring the relationship between image data in feature space. In [250], multiple instance learning is proposed for image ranking based on local similarity. Zhou *et al.* [354] proposed relevance feedback for retrieval. The framework combines semi-supervised learning with active learning to address the problem of limited training examples and biased label distribution. RankBoost proposed in [102], is subsequently applied to define an active learning based image ranking framework in [251]. In [211], Imbalanced RankBoost is proposed

for ranking over large-scale image collection. An experimental evaluation of different image ranking methods is presented in [92]. Keyword based querying is the most preferred interface for retrieval. The problem of relating the keyword to visual properties has been separately studied as annotation and tagging problem (Refer [329] and [97] for details). Once the annotation is done, query-specific learning can be performed for ranking. In this direction, Grangier and Bengio [120] proposed concept of margin maximization with retrieval performance for image ranking without requiring explicit annotation. In [282], generalized scenario of text based ranking i.e. multi-attribute query is addressed by solving multi-label classification problem. The recent work by Jain and Verma [144] proposed image re-ranking method by applying the Gaussian process based regression.

Different retrieval systems for the same query would retrieve different set of documents. The fusion intends to improve the overall retrieval performance by the aggregation of an independent set of retrieved documents. [101] presents a detailed survey of different fusion methods with information retrieval perspective. In the context of multimedia retrieval, independent retrieval may capitalize on single modality of the information. McDonald and Smeaton [207] presented an experimental evaluation of different fusion methods for video retrieval. The simplest approach in this direction is the merging of documents from different retrieval system and ranking the complete set [260, 259, 81, 83]. In [230], borda count model is used to combine the results obtained by annotation based search. Wei *et al.* [326] proposed cross-reference strategy for late fusion of multi-modal features for video indexing. In [178], multi-partite graph based clustering is proposed for image ranking which combines the individual cluster ranking with local ranking of every

image in a cluster by multi-objective optimization solution.

2.5 Concept driven Content Management

Multimedia content analysis by using semantic concepts requires procedures for extraction and modelling of latent semantics embedded in documents. Commonly, the problem of semantic concept learning is referred to as annotation which learns the correspondence between high level semantics with low-level features extracted at pixel, region or global level. The annotated document can be subsequently used for defining the indexing.

The simplest annotation model approach the problem as image retrieval problem, and does the annotation based on nearest neighbour based similarity. Another simple approach to perform image annotation is based on supervised learning [4, 315, 143]. Here each annotation defines a category, and a classifier trained with pre-annotated example images is used to label new image. The early works on concept based image annotations incorporated relevance feedback obtained from the user in different forms [328, 208]. Duygulu *et al.* [79] proposed generative modelling for region based image annotation. The presented translation model is shown to learn the underlying semantic concepts having similar visual appearance. In this context, Jeon et al. [147] presented cross-lingual information retrieval based cross-media relevance model (CMRM) for image annotation and ranked retrieval which showed better retrieval performance than [79]. The model defines a joint probability distribution of image regions and annotation words which is subsequently applied for generating the annotation of query image. Here, the regions are blobs generated after the

feature based clustering of image segments obtained by normalized cuts. Subsequently, Feng et al. [96] extended the concept to rectangular blocks and proposed continuous version of relevance model (MBRM) having multiple Bernoulli model for word probability estimates, and kernel density estimate for image feature probabilities. The conventional user prefers keyword based querying for image retrieval. The generative probabilistic models have shown good performance for document retrieval which are prominently text based query systems. These models define latent space representation between the document and word frequencies. The adaptation of such latent space models for image annotation have been demonstrated in [218, 344]. The bag-of-words model is used for image representation as well as training annotations. These models perform the unsupervised learning of occurrences between image features and annotation tags which is further utilized for image annotation. Yang *et al.* [346] applied multiple-instance learning for learning the correspondence between image regions and annotation keywords representing high-level concepts.

Some recent works have explored the combination of classifiers, and combination of features for image annotation. Xiaojun *et al.* [248] combined probabilistic output of two classifiers for annotation. The SVMs are used for classification, where the first SVM is trained on bag-of-words features, and the second SVM is trained on combination of color, texture, color histogram and edge histogram feature. In [324], combination of probabilistic models (RVM and CRFs) is explored for learning the high-level image concepts. Fan *et al.* [90] have proposed hierarchical classification for multi-level image annotation. The hierarchical classifier training is performed by boosting algorithm incorporating concept

ontology and multi-task learning. An extensive experimental comparison of some of the recent image annotation methods have been presented in [197]. The authors have shown that retrieval based on simple combination of different distance measure defined over low-level image features can provide effective annotation methodology. Here, the keyword assignment is performed by a greedy label transfer algorithm. TagProp [123] is another retrieval based image annotation model which combines nearest neighbour approach with distance metric learning in discriminative framework. In [179], the relevance learning based image tagging framework is presented which estimates the tag relevance by counting votes of the visual neighbours on tags.

In general, the lexicon of semantic concepts, define high-level concepts related to objects, scenes and events which could be conveniently learned by a simple annotator. For example, SVM based annotation to different concept classes. However, semantic concepts also present in the form of a set of intermediate concepts which require exploration of latent semantics of the content. In this direction, recent work has applied probabilistic modelling for concept learning for different applications. In [244] and [129] generative modelling is applied to video event recognition. The authors in [323] and [349] have proposed conditional random fields based methods for event detection. Also, the recent work by Shen *et al.* [278] have proposed a Latent Dirichlet Allocation based topic model for temporal event recognition. The use of intermediate concepts i.e. latent concepts for multimedia analysis provides an approach for multi-modal information fusion. In this direction, Barnard *et al.* [22] presented a probabilistic model based fusion of textual and image information. The model learns the joint model of image regions and associated

textual description by translation modelling using a hierarchical clustering based aspect model. The application of text description available with video for event recognition is explored in [341] and [340]. Blei and Jordan [30] learned the topic level conditional relationship between image regions and annotations by defining correspondence LDA. In [357], transductive learning is applied to model the semantic correlation between multi-modal data for indexing and retrieval. Messina and Montagnuolo presented cross-modal clustering based approach for multi-modal information fusion [214]. Rasiwasia *et al.* [254] presented cross-modal canonical correlation analysis for multi-modal retrieval using topic space representation of text modality and feature space representation of image modality. The recent work by Jia [148] relaxed the condition of one-to-one relationship between text and images as assumed in [30] to generalized scenario of image and textual descriptions available in loosely coupled fashion. Fundamentally, the model generalizes a Markov random field over the LDA topics derived from document collection. However, the Markov assumption does not account for large range dependencies between the documents. In this direction, Wu et al. [337] have presented most updated results by formulating the image tagging problem as matrix completion. Novel optimization framework using theory of composite function optimization is developed which has shown significant improvement in comparison with TagProp([123]) and method presented in [179].

2.6 Multimedia Retrieval and Analysis Systems

Section 2.1 to 2.5 discussed fundamental issues related with multimedia database indexing and related contributions. Initial work on multimedia retrieval experimented with the available methods primarily developed for text retrieval. For conventional text database management systems, inverted index is an indexing technique with sub-linear time complexity for locating an entry. Sivic and Zimmerman [286] have applied the technique for object retrieval by indexing the visual words learned for predefined visual dictionary. The technique removes the requirement for storage and comparison of high dimensional descriptors and provides access to the relevant frames. Subsequently, [339, 146, 150, 317] have applied the technique for large scale image indexing, and annotation applications. Tree based data structure are efficient solution for multi-dimensional indexing [266]. Early works in this direction have explored application R-tree for image indexing and retrieval [40, 88]. R-tree and its variant have been used most often for indexing high dimensional datasets. First generation of Query By Image Content (QBIC) proposed by IBM used R-tree as the underlying indexing technique [100]. Smith and Chang [290, 289] used quad-tree for indexing the large scale image collection for content based retrieval. In this context, Gong et al. [115] introduced the application of SR-tree for image indexing and retrieval. In general, KD-tree is most preferred for nearest neighbour searching in main memory [330]. Beis and Lowe [24] presented optimized k-d tree which applies modified search ranking so that bins in feature space are searched based on their closest distance from the query location. Silpa and Hartley [284] demonstrated the use of different randomiza-

tion techniques for improving the the performance of KD-tree based retrieval. White and Jain [330] proposed variant of KD-tree, and R-tree using KL transform, and demonstrated superiority of novel R-tree on the existing state-of-the-arts. Nister and Stewenius [227] proposed use of hierarchical k-means for learning the vocabulary tree of local descriptors. Philbin et al. [243], used approximate k-means for visual vocabulary learning based on approximate nearest neighbor method. The approach have showed significant improvement with comparison to hierarchical k-means. Here, the relevance score of an image to the query is obtained by inverted file where [227] have proposed hierarchical term frequency inverse document frequency score computed by the hierarchy of visual words. Early work in multimedia retrieval demonstrated the capability of text based methods. However, the growing complexity of the multimedia data in-terms of richness of information and amount of content have created the requirement of novel methods for tackling the problem. In the following discussion, we concentrate on domain specific issues and contributions for document images presented broad review existing methods.

2.6.1 Document Image Retrieval

Document images encompass broad range of documents having graphics/image embedded with text information in imaged form. The management of such documents requires significant amount preprocessing as noise removal, binarization and skew correction [91, 125, 268, 39]. A typical document image analysis framework proposed in [202] is represented in figure 2.1. The recognition requires the layout analysis for identifying different logical blocks in the page. The layout analysis first identifies the physical segments,

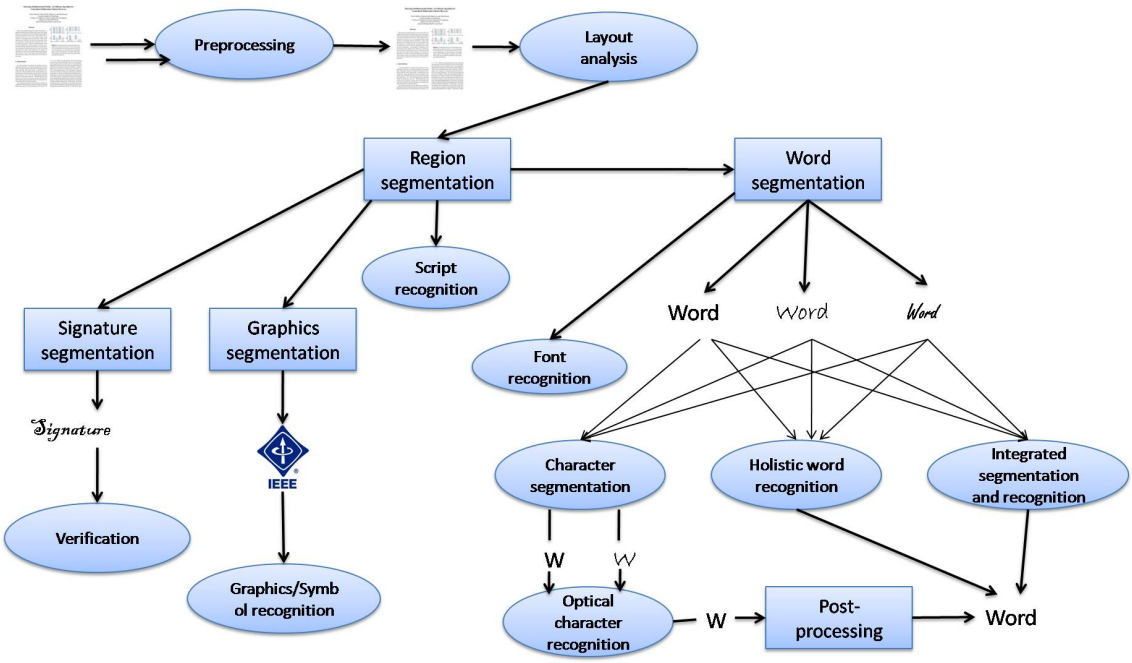


Figure 2.1: Generalized document image analysis flow-diagram

i.e., text, graphics and background regions. Further, the logical analysis of text region identifies the paragraphs, equations, signatures and tables [41, 11]. In [171], learning based logical labelling is applied for document image analysis. In [10], a comparative evaluation of recent page segmentation methods is presented on collection of historical document images. The knowledge of script of the textual content is the preprocessing requirement for efficient recognition. The automatic script identification is much researched problem. Some of the related contributions are present in [134, 335, 238, 273, 69]. The recognition converts the document image to text form which can be indexed and retrieved using existing text retrieval systems. However, the recognition does not always generate accurate text equivalent. In such a scenario, the indexing and retrieval of document images depends on image properties of textual regions. In this direction, some of the related work

is discussed in [20, 269, 205, 193, 204] A comprehensive survey of existing methods for printed document image retrieval is presented in [203]. Rath et al. [257] one of the earliest text based retrieval system for historical handwritten documents using relevance models. Novel set of feature description for handwritten word images is presented in [255]. Word spotting presents an efficient retrieval technique based on image matching based retrieval. Manmatha et al. [198] introduced the concept for text based handwritten document retrieval using simple image matching concept. Further improvements in this direction have been reported in [258, 103, 104, 99]. The application of Hidden Markov Model for developing script independent word spotting model have shown in [263, 336]. One of the earliest work in handwritten document recognition by Kim and Govindraju introduced chain code based handwritten word recognition in [153]. Subsequently Madhavnath and Govindraju [195] developed holistic paradigm for handwritten recognition by considering word image as single entity. In particular, Bunke et al. have produced significant contribution towards the development of graphical modelling based methods for handwritten document recognition [35, 140, 36, 358]. In this direction, recent work presented in [99] developed lexicon free keyword spotting for handwritten documents using character level trained hidden Markov models. With reference to Indian scripts, the state-of-the-art of recognition and retrieval technologies for Indian script documents has not matured yet. Some of the earlier works for printed Indian script document image retrieval have been discussed in [49, 267, 213, 160, 275]. Recent work by Krishnan et al. [157] have proposed fusion of recognized document and image based representation for Indian script document retrieval. The recognition and retrieval technology for Latin script printed documents

have reached upto the state-of-the-art level with several commercial products are widely available [303, 17, 16, 305, 304]. Text detection from images and videos have attracted significant research interest for various applications e.g. searching, storage and summarization. Earlier works in this direction are reported in [177, 87, 348, 187, 107]. In [118], novel set of features based on discrete cosine transform have been proposed for locating the textual regions in natural scenes. Novel application of Random Ferns for character detection in scene images is demonstrated in [322]. The recent works by Shivakumara et al. proposed efficient detection frameworks for addressing the multi-oriented text appearance in videos [281, 280].

2.6.2 Video Analysis and Retrieval

A typical video content analysis and retrieval framework requires four primary processes: feature extraction, structure analysis, abstraction and indexing [75]. The video contains

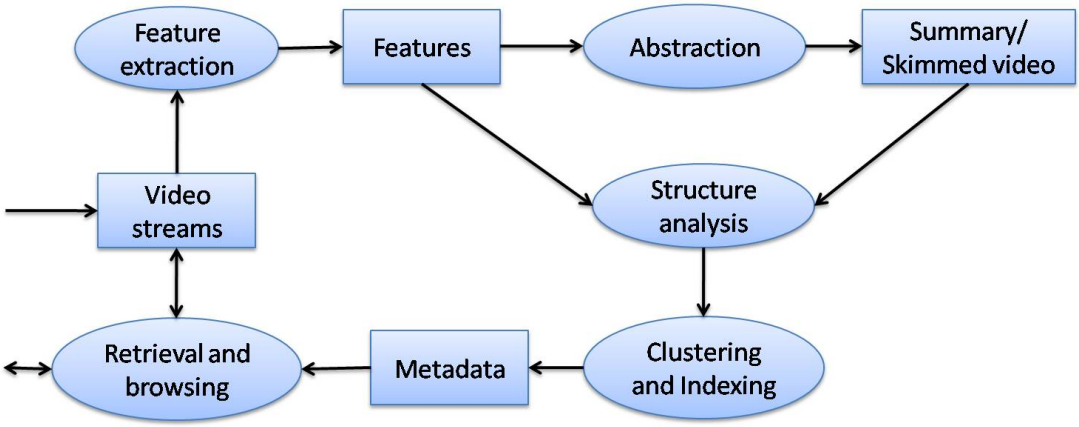


Figure 2.2: Generalized video content analysis and retrieval flow-diagram

temporal information associated with the sequence of video frames. The temporal infor-

mation results in structure and motion features in video content. Recent research has been done on spatio temporal descriptors [152, 76, 332, 169]. The structure analysis involves detecting temporal boundaries and identifying meaningful segments from video such as shots and scenes. The structure information can be applied for recognition, retrieval and summary generation [262]. The review of earlier video segmentation methods are discussed in [59]. In [45], spectral clustering is applied for scene segmentation from video. The summary of existing shot segmentation methods are presented in [287]. The recent work by Sidiropoulos *et al.* [283] presented multi-modal approach for scene detection by probabilistic merging of semantic visual concepts and audio events. The recent development multimedia applications have motivated significant contribution towards event detection and activity recognition from video streams [137, 210, 355].

In general, the audio information for multimedia analysis has been exploited in combinatorial fashion with video information. The audio information is primarily described using acoustic based based features, e.g., pitch, loudness, bandwidth and harmonics [333, 133]. Mel-frequency cepstral coefficients (MFCC[212]) has been the preferred audio feature representation approach. In [124], the MFCC with set of perceptual features have been used for SVM based audio retrieval. The OWL ontology based classification and retrieval of music documents is presented by Ferrara *et al.* [98]. Chechik *et al.* [51] have presented text based audio retrieval using Gaussian mixture models (GMM) and SVM.

2.7 Motivation for the Present Work

Sections 2.1 presented existing methods for using multiple features for recognition and retrieval. The section 2.2 briefly discussed different methods for multimedia indexing and retrieval. We have explored the advancement in nearest neighbour based retrieval in section 2.3. Subsequently, learning based methods for content and concept based multimedia retrieval. Finally, a brief discussion on document specific multimedia analysis and retrieval methods are discussed in section 2.6. Based on the survey, we identify the following key-points which have motivated for the work presented in this thesis.

- Multiple features have been extensively applied for recognition problems. Nevertheless, multiple feature based recognition having a large number of categories is not attempted. Recognition problems have established MKL theory which can efficiently combine multiple feature based document representation for classification. The retrieval problems have not explored principled methodology for combining multiple features.
- The hashing based approximate nearest neighbour search presents robust retrieval approach for high-dimensional multimedia data. The effectiveness of hashing based indexing can be improved by similarity based grouping. However, learning the hashing using distance based grouping is not yet explored for multimedia indexing.
- The existing multimedia management methods for image form of text documents have not addressed the requirement of old and degraded documents having complex layouts and multi-modal information content.

- The effectiveness of concept based multimedia analysis depends on identification accuracy of high-level semantic concepts from the document. In this context, multiple feature based semantic concept learning is not explored. The multi-modal information associated with document improves the understanding of inherent semantics. The probabilistic learning for modelling the multi-modal context at concept level and feature level has not been explored.

Motivated by these observations discussed above, we primarily address the following two problems.

- Feature combination for recognition/retrieval.
- Multi-modal fusion for retrieval at feature and concept level.

The first problem concentrates on documents having textual information in image form. The second problem deals with general class of documents having text and graphics information and pure images. In this direction, following are contributions of this thesis:

- A robust classification framework by employing multiple feature based representations for character/symbols images is proposed. The framework applies MKL for the optimal combination of different feature representations. In this direction, the work also makes a novel attempt to explore learning based approach for feature combination in Indian script optical character recognition.
- A novel feature representation for binary patterns is proposed. The feature uses a novel grid based approach for shape information extraction which does not require edge, contour or skeleton detection. The approach implicitly handles general

scenarios and can distinguish objects having similar outer envelopes but different inner contours. Also, the proposed feature represents the objects as constant dimensional vectors which gives flexibility to use direct comparison methods for similarity matching.

- The use of DBH for document indexing is proposed. A new approach for selection of pivot objects for hashing functions has been explored for increasing the collision probability. A hierarchical scheme for organization of hashing tables have been presented. Novel framework for multi-probing in binary mapping function is presented.
- Existing multiple feature based indexing schemes primarily follow heuristic based approach. We propose a novel indexing and retrieval framework which unifies the learning based feature combination methodology with approximate nearest neighbour search. The kernel distance based hashing is proposed to accommodate kernel matrix based distance measure. A new MKL formulation for retrieval is proposed which learns the optimal kernel for indexing as a parametrised linear combination of supplied kernels. A novel application of Genetic algorithm for solving the optimization problem is proposed. The weighted linear scheme identifies an optimal combination of base kernel matrices by solving the optimization problem using a Genetic algorithm.
- Novel methods for multi-modal information integration from document images are proposed. We propose text/graphics segmentation method for documents having complex layout. The framework exploits color and texture property for identifying

different segments in color planes. The contextual dependency across the color planes and spatial proximity is modelled for final labelling. The application of the framework is subsequently demonstrated for multi-modal retrieval of document images. We propose a novel method for performing computationally fast script identification in documents having random usage of different scripts.

- We propose a novel method to use the recognizer's characteristics for efficient indexing and retrieval of documents having recognition errors. Topic modelling based retrieval framework for noisy text documents is proposed by applying LDA. A methodology to improve the word based retrieval of noisy text documents is proposed which uses noise invariant LDA topic distributions with shape characteristics of corresponding word image.
- We propose MKL based concept learning framework for recognition and retrieval which uses multiple feature representation for improved bridging between the low-level features and high level semantic concepts. A new probabilistic learning based framework is proposed for learning the correspondence between the multi-modal concepts for retrieval. The use of probabilistic learning is subsequently applied for defining semantic classification framework for multi-modal documents.

Chapter 3

Multiple Features for Recognition of Binary Patterns

3.1 Introduction

Binary pattern classification is an important problem in various real-world recognition applications. These applications require two separate modules; an effective scheme for a representation of the pattern in terms of useful features and a robust classification system. The effectiveness of a feature is characterised by its invariance to geometric transformations and noise for samples belonging to the same class, and by high variance across different classes. The classifier is expected to learn the discriminating boundary between known categories for robust classification. Conventionally color, shape and texture have been the preferred cues for object representation in computer vision. However, shape and structure are the primary attributes for feature based representation of binary pat-

terns. Ability of a feature to capture distinctive characteristics of classes vary with the inherent variabilities embedded in the individual classes. The use of different feature sets can provide complementary information about characteristics of different classes. The recognition performance can significantly increase by a principled combination of different features. Combinations of features for recognition is a well researched problem [296, 359, 297, 142, 353, 80, 270]. In this chapter, we explore the application of multiple features for binary pattern recognition. Additionally, a novel feature representation for binary patterns by utilizing the object shape information is proposed. Two important binary pattern recognition problems namely character primitive recognition and symbol recognition have been addressed in this work. These problems offer similar challenges for classification in terms of large categories and variation in examples. These variations include non-linear transformations and distortions including elastic and non-elastic deformations.

Recognition of primitives is an important step in optical character recognition (OCR). The basic OCR system segments a word image in to a set of character/symbol primitives and recognizes them using a classifier trained over standard templates. Indian scripts belong to alpha-syllabic writing system with the basic unit consisting of consonants and vowels. The vowels are practised in independent/dependent manner. The use of dependent vowels and modifiers (*maatraas*, diacritic marks etc.) with base consonants exhibits large variations. Additionally, a large set of compound characters is generated by consonant combinations (half and full forms), vowels and modifiers feature in the script. The segmentation in such a scenario generates a large primitive set of character glyphs and base

characters. The recognition consequently results in a large category classification problem. Simultaneously, these primitives exhibit complex combinations of cursive, horizontal and vertical structures. The recognition performance therefore depends on the feature space definition as well as generalization ability of the classifier. In such a scenario, the application of multiple types of features exploiting inner detailed structural characteristics as well as the envelope shape of the primitive can provide the desired classification accuracy.

The symbol recognition problem cover a range of identification tasks performed in applications like document image analysis, graphics recognition, trademark/logo retrieval. In this chapter, we also deal with the problem of the recognition of binary graphic patterns e.g. logos, trademarks or silhouette images as symbols. Symbols are used for a graphical mode of information sharing and usually appear in isolation. The appearance of symbols undergoes changes due to view point variations, geometric transformations and shape distortions. Symbols also undergo morphological distortion such as protrusions, incision or hollowing in silhouettes. The visual properties of symbols also vary widely with changes in application scenarios. The major contributions in this chapter are as follows.

- We propose a learning based framework to optimally combine features in large category problems. The framework exploits the existing theory of Multiple Kernel Learning (MKL), which has shown excellent performance in many recognition applications [114, 312, 167]. Fast classification is an essential requirement for practical recognition applications. A novel multi-class framework employing MKL in directed acyclic graph (DAG) architecture is presented for fast classification.

- A novel shape based feature representation for binary patterns is proposed. The representation provides rich description of the pattern by distance and orientation based distribution of pairwise arrangement of sampled object boundary points.

The character classification experiments concentrates on two prominent Indian scripts namely Gujarati and Bengali. The symbol recognition experiments use the MPEG-7 shape dataset available at [302]. The following section presents review of existing work in Gujarati and Bengali OCR and symbol recognition in general. Subsequently, the novel shape descriptor is introduced in section 3.3.3 with the details of other features.

3.2 Related Works

This section provides a review of existing work in the field of Indian script OCR and and symbol recognition. The character and glyph primitives across Indian scripts exhibit significant variations. However, the present work focusses on Gujarati and Bengali script characters for evaluation and analysis. In addition, the recent work in symbol/logo recognition is also discussed.

Gujarati is a prominent Indian script used by approximately 60 million people. The script has 12 vowels and 34+2[1] consonants. In addition to basic symbols, 'vowel modifiers' (total 12 symbols) are used to denote the attachment of vowels with the core consonants. Consonant-Vowel combinations are defined by attaching an unique symbol for each vowel to the consonant, called a dependent vowel modifier. The dependent vowel can appear

[1]Two conjuncts /ksha/ and /jya/ are also treated as a basic consonant

before, after, above, or below the core consonant. In addition to the basic consonants, Gujarati also uses consonant clusters (conjuncts). The development of Gujarati OCR is in its infancy compared to other Indian scripts such as Devanagari, Bengali, Telugu and Tamil. The earliest work on Gujarati character recognition reported by Antani and Agnihotri applied 1^{st} and 2^{nd} order *Hu* moments for character image representation [9]. The application of these features for recognition is demonstrated using Nearest Neighbour (NN) and Minimum Hamming Distance based classification. Dholakia *et al.* [72] presented zone wise identification algorithm for Gujarati script which separates the text into three logical zones i.e. lower and upper modifier, and middle character. The identification helps to segment the character/symbols at zone level (Figure 3.1). The strategy generates a finite number of symbol categories therefore resulting in less misclassification. Following the similar approach in [73], the authors applied *Daubechies D4* wavelet transform based feature representation. For the subset of middle zone symbols having 119 categories and 4173 examples (excluding the dependent vowel modifiers), the authors reported a recognition accuracy of 97.59% with a Neural Network based classifier. The absence of shirorekha

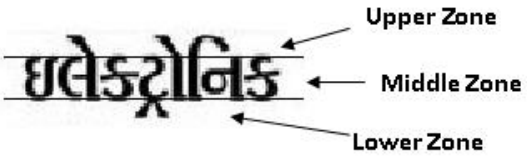


Figure 3.1: Zoning in Gujarati text

(horizontal line at the upper part of word formation) in Gujarati script means that it displays a unique appearance compared to Devanagari and Bengali. The recognition of indepen-

dent symbols from each zone requires the identification of symbols as *middle zone symbols* (consonants + conjuncts + vowel modifier corresponding to /AA + half forms of consonants + half form of conjuncts), *upper zone symbols* (upper parts of vowel modifiers) and *lower zone symbols*. The possibility of distinct symbols defined by this combination is significantly higher than 119. The majority of these symbols are from the middle zone having complex shapes e.g. combination of consonants, half form of consonants and conjuncts. Additionally, some symbol categories are very similar in appearance; therefore, making the recognition more difficult. Considering the large symbol set and shape complexity, the problem of Gujarati character recognition is still a challenging problem. We propose the application of multiple features for recognition because single feature representation is not sufficiently discriminative. The novel concept of multiple kernel learning (MKL) provides an efficient solution here. The MKL algorithms learn an optimal combination of set of kernels as part of its training, therefore providing an intelligent method to combine multiple features. It also provides the flexibility of learning the best kernel for the task compared to the traditional method of cross-validation.

Bengali is a widely researched script after Devanagari among Indian scripts for OCR development. The script consists of 11 vowel and 39 consonant characters in basic set. The shapes of vowels are modified when used with consonants except when used in the start of word or in pairs. Additionally, two or three consonants may be combined into complex shapes. These compound characters may appear alone or attached with a vowel modifier. The characters in a word are usually connected through headline (Shiro-rekha or Matra). Therefore, the OCR processing is required to segment/identify individual char-

acters. The earliest work on Bengali OCR reported in [261] used NN based classifier and connected components based features for recognition. In [235], Pal and Chaudhuri presented an OCR for uni-font Bengali documents. The recognition follows a zone-wise approach, which first identifies three prominent horizontal zones in text followed by a linear scanning based character segmentation step. A multi-stage recognition framework is presented by applying structural and shape features over tree based classification. The details of character segmentation in Bengali OCR are discussed in [109]. Following the zone separation approach, Sural and Das [298] presented a hierarchical framework employing Multi Layer Perceptron for recognition. The authors proposed Hough transform based fuzzy features for addressing the noise presence in scanned documents. In [196], the author applied a curvelet transform in combination with NN classifier for recognition. However, the performance of this approach on noisy and degraded documents is not documented. The recent work by Pal el al. [239] presented an OCR for complex documents printed in different styles. However, these methods exploit the information represented through a single feature set. The performance of such frameworks is not guaranteed in the presence of font variations as a single feature space based representation does not guarantee equal discrimination among all category of examples exploiting envelope and inner structural detail based features. Additionally, the existing work in Bengali OCR have not significantly exploited the maximum margin based classification (SVM) widely regarded as the state-of-the-art. In this case, our work presents a robust classification framework by employing multiple feature based representations for character/symbols images. The work also makes a novel attempt to explore the learning based approach for feature combination

in Bengali OCR.

Feature Representations for Symbol Image

Symbol recognition in the recent past has received significant research attention because of its importance in various applications. In this context, a large contribution has been made in both feature representation and classifier architecture. The comprehensive survey of early related work is presented in [191, 53]. Feature extraction algorithms have primarily exploited shape information for symbol description. Introduction to various shape descriptors following MPEG-7 standards is presented in [32]. In this context, Mokhtarian *et al.* [1] defined a curvature scale space for object image using the outer boundary information of closed contours. In [29], a symbol description is generated using a representative set of selected from the training examples. Alajlan *et al.* [6] presented a triangle area representation by estimating the areas of triangles within the closed object boundary. The representation is scale, translation and rotation invariant. However the invariance to deformations is not established. The shape context [25] based symbol description represents global object shape characteristics by a set of descriptors computed over points sampled along the boundary. The representation is robust to occlusions and deformations. However classification requires correspondence matching for similarity based labelling. Escalara *et al.* [85], presented a blurred shape model for symbol representation to address soft, rigid and elastic deformations. The blurred shape model uses the spatial probability of the appearance of shape pixels and their contexts for descriptor computation. The extension for rotational invariance in the model is presented as a circularly blurred shape model [84].

However the robustness of these feature representations in the case of symbol rotations, partial occlusions and deformations, and high intra-class variations is not guaranteed. We propose a new feature definition to address the problem of deformations which is less sensitive to occlusion and invariant to rotation. Also, a learning based feature combination for developing the symbol recognition engine is proposed.

3.3 Feature Extraction

The details of the feature sets used in this work are as follows.

1. Fringe Map: a distance transform based feature representation technique
2. Histogram of Oriented Gradients
3. Shape Descriptor: a shape context based global shape descriptor
4. Modified Shape Descriptor: a modified version of shape descriptor to take care of shape deformations

The feature representation details and computation process are presented in the following discussion.

3.3.1 Fringe Map (FM)

The Fringe map essentially represents the distance transform of binary symbol/character image. The transform extracts the inter-pixel distance relationship in the image, and presents the knowledge as a distance map [264]. In this sense, for a binary image having a set of feature (foreground) and non-feature pixels background pixels), the distance map

defines its gray scale equivalent. The computation process of a distance map for given binary image converts the positions of non-feature pixels by their distances to the nearest feature pixel. The positions of feature pixels are replaced by $0's$. E.g., for a given binary image $I(x, y)$, let $F_{feature} = \{(x, y) : I(x, y) = 0\}$ be the feature pixel set and remaining pixels $F_{nonfeature}$ are back-ground pixels. Mathematically, the distance transform is defined as:

$$D(I(x_i, y_i)) \equiv \inf\{\| \ (x_i - x, y_i - y) \ \|\}_p : (x, y) \in F_{nonfeature}\}$$

Here $\| \cdot \|_p$ is the L_p norm metric applied for computation. In the present work, Manhattan distance is considered such that generated distance maps consists of integer values. The conventional distance transform consider all eight directions for map computation; however, the present work has considered only four prominent directions to reduce the transform computation time. In this sense, our distance map is the approximate form of distance transform which considers only the four nearest neighbours. The fringe map computation reads the original binary image and marks the feature pixels as $0's$. The scanning process leaves $0's$ unchanged whereas $1's$ which are $4-neighbors$ of $0's$ also remain unchanged. These $1's$ are marked, and the subsequent computation converts remaining $1's$ which are $4-neighbors$ of marked $1's$ to $2's$ and identifies them as marked. The procedure follows until all the non-feature pixels are marked. Figure 3.2 shows the Fringe map computation steps for sample Bengali character image. The preprocessing step for transform computation includes bound box detection of the symbol image. The FM computation is parameter independent process; therefore, simplifies the parameter selection problem.

Additionally, the feature representation is of constant dimension equivalent to normalised size of example image which can be conveniently used with different classifiers.

3.3.2 Histogram of Oriented Gradients (HOG)

Histogram of oriented gradients is a directional feature which is suitable for binary and gray scale image representation [62]. It has been popularly employed in many computer vision and character recognition problems. Earlier, Favata et al. [95] introduced a similar feature extraction method using gradient information for binary symbol representation. The HOG feature represents the distribution of orientation gradient in the local neighbourhood. In this case, local gradient is computed by dividing the image in smaller regions defined as *cells*. In [62], different smoothing and derivative operators were experimented for gradient computation. They achieved best results with gradient computed on unsmoothed image with directional derivative computed with simple 1-dimensional operator $[-1, 0, 1]$. We follow this strategy for local gradient computation. In the present work, the contribution by each pixel in the local histogram computation is weighted by gradient magnitude at the pixel location. The illumination and contrast invariance is obtained by the gradient normalization in local region. The normalization is performed by defining larger regions *blocks* by combining neighbouring *cells*. *Blocks* may be overlapping or non-overlapping, however we have considered non-overlapping blocks for normalization. The rectangular *blocks* are defined by grouping adjacent *cells*, and L_2 norm of the concatenated histograms corresponding to *cells* is computed for obtaining the normalization factor.

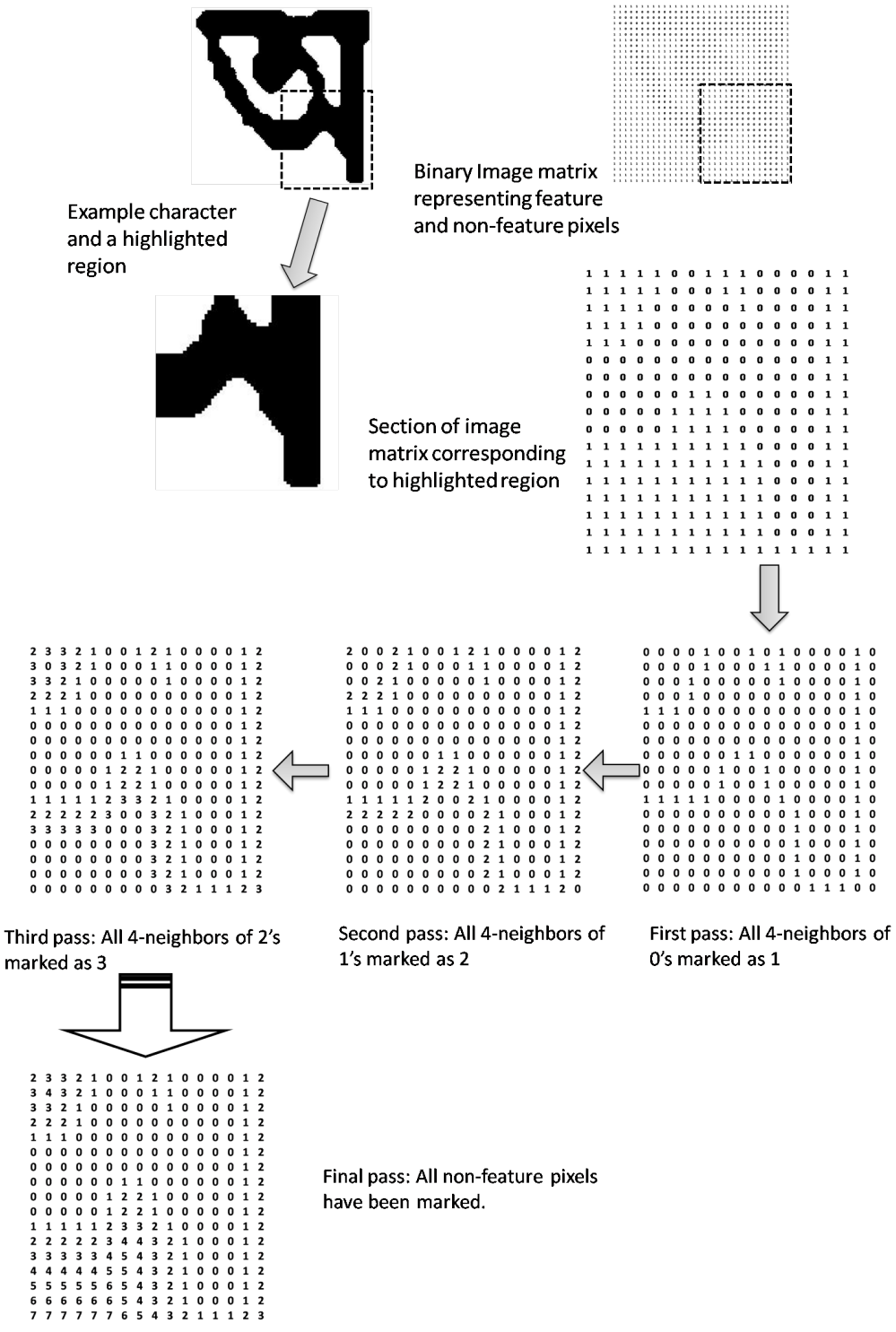


Figure 3.2: Fringe map computation steps for example character image

3.3.3 Shape Descriptor (SD)

The following discussion introduces a novel shape descriptor which defines a generic representation scheme applicable for different binary patterns including character, symbol and word images. Shape information is an important visual cue for object recognition. In the available literature, various image descriptors utilizing the object shape information have been proposed [189, 288, 216, 121]. These descriptors represent the object shape information in various forms including chain code, polygonal approximations, transforms, curvature and moments. Broadly these methods are categorized as contour based and region based shape descriptors. The contour based shape descriptors make use of only the boundary information of shape extracted by a conventional edge or contour detection approach. The region based shape descriptors exploit the pixel intensity information within the shape region.

We propose a contour based shape descriptor which also employs the characteristics of the inner contours of complex shapes for descriptor computation. The information of inner contours gives distinct representation to objects having a similar outer boundary e.g., Devanagari character pairs {/sha, /pa} and {/ba, /va} in Figure 5.6. The conventional

Figure 3.3: Example character objects

approach of edge or contour detection for boundary extraction is highly noise sensitive. Therefore, the limitations of edge detection based boundary extraction is addressed by

following a novel grid based approach. The computation process for feature representation is divided into two steps:

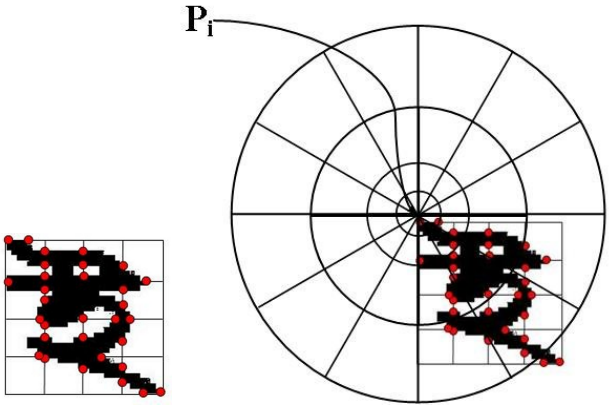
i Descriptor point extraction for shape representation.

ii Feature computation from the set of descriptor points.

Descriptor Point Extraction

The conventional approach to extract points for boundary based descriptor is by detecting shape contour and randomly sampling the points from the contour coordinate set. The object's shape is represented by a point set $P = \{P_i\}$ for $\{i = 1, ..., l\}$. Subsequently, the set P is used for shape descriptor computation. The contour extraction approaches have inherent limitations because of the involvement of edge detection. In addition, the random sampling does not guarantee a uniform distribution of points on the boundary. In our approach, we convert the gray scale image to a binary image by a standard binarisation routine. The preprocessing step include object bounding box detection and normalization of the bounded image by aspect-ratio preserving scaling transform. A logical grid of constant size is placed over the normalized image, where the transition points obtained by traversing on the grid lines define descriptor point set P. In addition, points lying on the boundary and coinciding with grid lines are also selected as descriptor points. A transition point is marked on the grid in case of intensity change, i.e., 0→1 or 1→0. The grid based approach for descriptor point extraction gives better distribution with respect to distance and orientation for complex shapes having multiple inner contours. In addition, the shape information embedded in inner contours of complex shapes can also be used efficiently

in this scheme. Figure 3.4a shows the red dots as transition points on the horizontal and vertical grid lines.



(a) (b)

Figure 3.4: Descriptor points on character image and point P_i in log-polar space

Shape Representation

The density of shape descriptor points varies with the complexity of object shape in the image. The distribution of these points based on relative arrangement is represented by log polar histogram. For a point set $P = \{P_i\}$ for $i = 1, ..., l$, the log polar histogram is defined as

$$H_i(k) = [q \neq p_i | (q - p_i < bin(k))]$$

Here, $H_i(k)$ is the shape context of point P_i with each bin k representing the position based count of other descriptor points with log polar centred at P_i. The bins are uniform in log-polar space, incorporating sensitivity for neighbouring points. Let D be the distance, and

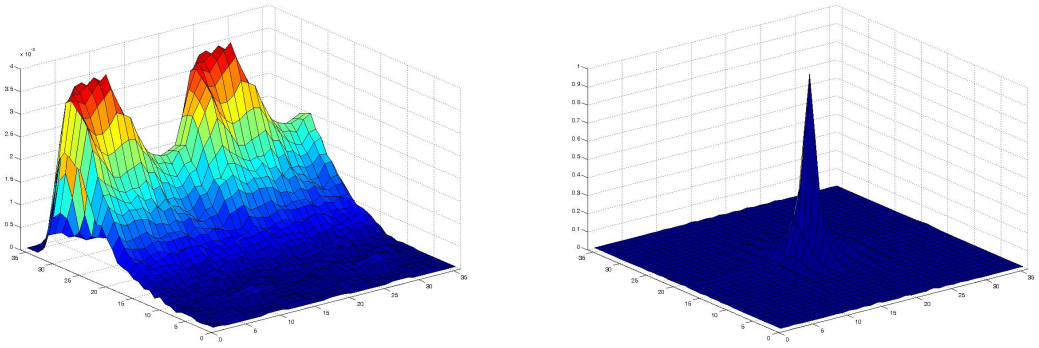
A be the angle matrix for point set P. Let $[d_0, d_1] \cup [d_1, d_2] \cdots \cup [d_{m-1}, d_m]$ are distance, and $[\alpha_0, \alpha_1] \cup [\alpha_1, \alpha_2] \cdots \cup [\alpha_{n-1}, \alpha_n]$ are angular bins. Here m and n are number of distance and angular bins used for histogram computation. For point P_i, the $bin(p, q)$ of h_i is defined as

$$h_i(p, q) = \sum_{j=1, j \neq i}^{l} \delta(D_{i,j}, A_{i,j}, d_{p-1}, d_p, \alpha_{q-1}, \alpha_q),$$

Here

$$\delta(D_{ij}, A_{i,j}, d_{p-1}, d_p, \alpha_{q-1}, \alpha_q)$$
$$= \begin{cases} 1 & \text{if } \ D_{i,j} \in [d_{p-1}, d_p],\ A_{i,j} \in [\alpha_{q-1}, \alpha_q] \\ 0 & \text{otherwise} \end{cases}$$

The relative arrangement of descriptor points is unique for each word shape. Integration of shape context for all the descriptor points describes the global distribution of the point-pairs n log polar space. Let h_{sum} is the integration of all shape contexts, i.e., h_{sum} is the cumulative histogram that represents distribution of log distances and orientations between the points in P. In h_{sum}, each $bin(p, q)$ represents the count of points, which are relatively arranged within distance $[d_{p-1}, d_p]$ and orientation $[\alpha_{q-1}, \alpha_q]$. We name the normalized h_{sum} as point distribution histogram (pdh). The point distribution histogram represents the structural arrangement of a shape and provides easy access to inherent semantic information embedded in the object shape. Similar shapes give rise to similar point distribution histograms. The local shape properties represented by closely positioned point-pairs are identified by the rising gradient of the normalized h_{sum} as seen from the front in the Figure 3.5a.



(a) Normalized h_{sum} (b) Fourier coefficients

Figure 3.5: Normalized h_{sum} and absolute Fourier coefficients of h_{sum} for the image in Figure 3.4a

The global shape properties represented by distantly positioned point-pairs are identified by the peak and falling surface of the normalized h_{sum}. The normalized h_{sum} is treated as image and we take its Fourier transform. The Fourier transform represents the characteristic function of point distribution histogram and captures the periodicity in h_{sum}. The amplitude information of Fourier coefficients define the shape descriptor for object feature representation (Figure 3.5b). The feature computation steps are summarized as:

- Extraction of point set P and computation of shape context h_i for each point in P
- Computation of h_{sum} as $\sum_{i=1}^{l} h_i$
- Shape descriptor $F(P)$ is defined by the magnitude of Fourier transform of *normalized*(h_{sum})

3.3.4 Modified Shape Descriptor (MSD)

The shape descriptor presented in section 3.3.3 conveys the object shape information by exploiting the distribution of relative distances and orientations of boundary points. The boundary points are the set of descriptor points which are sampled from the inner and outer contour of the object. The distribution referred as a *point distribution histogram* is equivalently the distribution of point-pairs based on their structural arrangements. The robustness of the descriptor is established by the dense distribution of point-pairs which are invariant in the case of small shape variations and deformations.

Nevertheless, in practice many symbols appear in a distorted form having protrusions, incision or elastic deformations. These distortions generate noisy point-pair distributions. The following discussion presents a novel extension to ensure the robustness of the shape descriptor in such situations. We consider a smoothed set of descriptor points in local neighbourhood. The operation results in a blurred version of descriptor point set. The new point set is less sensitive to the artificial distortions and encodes the global object shape information. The spatial neighbourhood for smoothing the descriptor point set P is defined by placing a logical grid S_{centre}. The grid S_{centre} is smaller or equal to the original grid S used for extraction of P. The new point set P_{centre} is defined as the centroid of the descriptor points from P located in the rectangular regions of grid S_{centre}. The centroid computation is done by averaging the local points in a region. Figure 3.6 shows an example symbol image. Grid S is shown by continuous black lines and S_{centre} is shown by dashed blue lines. The red circles are the descriptor points forming set P as the transition points

over S. The green plus signs are P_{centre} obtained as centroid of points in the rectangular region defined by logical grid S_{centre}.

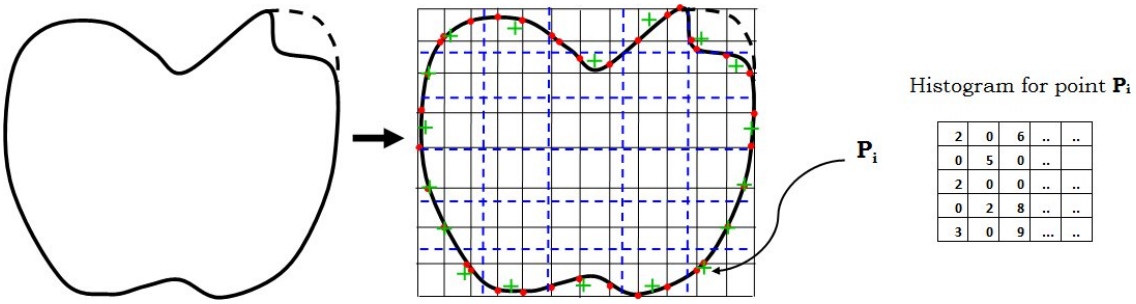


Figure 3.6: Modified shape descriptor computation

Subsequently, the modified shape descriptor is computed using the point set P_{centre}. The computation steps follow the procedure as discussed in section 3.3.3 with P_{centre} replacing point set P. The use of P_{centre} gives a robust feature description to the symbols in the case of artificial shape deformations. In addition, the computationally complexity of the modified shape descriptor is significantly less because $|P_{centre}| < |P|$.

3.4 Multiple Kernel Learning for Character/Symbol Classification

In general, the classification problem addressing character recognition and symbol recognition is a large category problem. Fundamentally, OCR processing is done using two strategies. The first strategy follows direct recognition of the complete consonant-vowel, consonant-vowel modifiers or conjunct-vowel combination. The combinations generate

a large number of primitive classes, the recognition of which is practically a very challenging task. In the second strategy, we can segment the consonants from the dependent vowel modifiers and recognize them separately. In this strategy, the primitive categories are significantly reduced. For example, extending the second strategy to zone wise recognition of Gujarati script reduces the primitive categories to 10%. In this case, classification complexity is also reduced by the utilization of zone information as the process requires three separate classifiers for the task. Additionally, character segmentation following the zone based separation generates a highly dense distribution of primitives in middle zone whereas sparse primitive distribution in the upper and lower zones. This simplifies the recognition problem as a simple classifier e.g., a minimum distance classifier, or a linear discriminant or template based matching can perform efficiently for lower and upper zone primitives. Nevertheless, the total number of primitive categories following zone-wise or direct recognition of all primitives is still large primarily because of minor inter-class variations between many primitives. For example the authors considered 119 symbol classes for Gujarati OCR in [73], and the Bengali OCR discussed in [46] considered 300 symbol categories. Also the frequency distribution in this case is highly unbalanced due to the rare occurrence of many characters.

Symbols mostly appear in an isolated fashion. The selection of the symbol set is generally context and application specific. Nevertheless irrespective of the application domain, large sets of conventionally significant symbols is prevalent in regular use. The appearances are generally influenced by rigid and non-rigid transformations for beautification resulting in high within class variation. Large symbol categories contributing high

between class variation and complex within class variations present a challenging task in designing an efficient symbol recognition system. In this section, we present a novel MKL framework for large category recognition problems. We concisely review existing MKL formulations and introduce the binary MKL adopted in this work. The subsequent discussion presents the details of the proposed DAG based architecture multi-class MKL.

3.4.1 Binary MKL Problem Formulation

We have adopted the MKL-SVM formulation proposed by Rakotomamonjy *et al.* [252]. The formulation presents an efficient approach to learn a sparse combination of kernels thus making it applicable for large scale problems. The sparsity of the linear combination of kernels is controlled by a L_1 norm constraints on the kernel weights. The decision function of a kernel based SVM for an input x is defined as $y(x) = f(x) + b$, where $f \in \mathcal{H}$. The $\mathcal{H}$ is RKHS associated with the $best$ kernel K. If we prefer to use a combination of kernels instead of using the $best$ kernel, the above decision function is modified as $y(x) = \sum_k f_k(x) + b$, where $f_k \in \mathcal{H}_k$. The $\mathcal{H}_k$ is the RKHS associated with kernel K_k. Let us consider $\boldsymbol{\eta}$ as the vector representing the kernel combination weights. The formulation of the primal optimization problem of MKL-SVM is done by incorporating a weighted L_2 norm regularization:

$$\begin{array}{ll} \min_{f_k,b,\zeta,\eta_k} & \frac{1}{2}\sum_k \frac{1}{\eta_k}\|f_k\|^2 + C\sum_i \zeta_i \\ \text{such that} & t_i \sum_k f_k(x_i) + t_i b \geq 1 - \zeta_i \quad \forall i \\ & \sum_k \eta_k = 1, \eta_k \geq 0, \zeta_i \geq 0 \quad \forall k, \forall i \end{array} \tag{3.4.1}$$

The above optimization problem can be decomposed into two steps. In the first step the f_k, b and ζ_i are learned with fixed $\boldsymbol{\eta}$. In the second step $\boldsymbol{\eta}$ is optimized through a descent step towards the minimum of the $J(\boldsymbol{\eta})$. The two steps can be represented as

$$\min \quad J(\boldsymbol{\eta}) \quad \text{such that} \quad \sum_{k=1}^{M} \eta_k = 1, \quad \forall \eta_k \geq 1 \quad \text{where} \tag{3.4.2}$$

$$J(\boldsymbol{\eta}) = \begin{cases} \min_{f_k, b, \boldsymbol{\zeta}, \eta_k} & \frac{1}{2} \sum_k \frac{1}{\eta_k} \|f_k\|^2 + C \sum_i \zeta_i \\ \text{such that} & t_i \sum_k f_k(x_i) + t_i b \geq 1 - \zeta_i \quad \forall i \\ & \zeta_i \geq 0 \qquad\qquad\qquad\qquad \forall i \end{cases} \tag{3.4.3}$$

The constraints of the optimization problem defined in equation (3.4.2) are over the simplex. The problem is minimized by a reduced gradient method, assuming that $J(\boldsymbol{\eta})$ is differentiable. Once the gradient of $J(\boldsymbol{\eta})$ is computed, $\boldsymbol{\eta}$ is updated in the descent direction such that the constraint on the simplex as well the positivity constraint of $\boldsymbol{\eta}$ are satisfied. The smoothness of f_k is controlled by η_k. The dual of the convex optimization problem (3.4.3) is defined as

$$\begin{aligned} \max_\alpha \quad & -\tfrac{1}{2} \textstyle\sum_{i,j} \alpha_i \alpha_j t_i t_j \sum_k \eta_k K_k(x_i, x_j) + \sum_i \alpha_i \\ \text{such that} \quad & \textstyle\sum_i t_i \alpha_i = 0 \\ & C \geq \alpha_i \geq 0, \ \forall i \end{aligned} \tag{3.4.4}$$

Thus following the strong duality the objective function is redefined as

$$J(\boldsymbol{\eta}) = -\frac{1}{2} \sum_{i,j} \alpha_i^* \alpha_j^* \sum_k \eta_k K(x_i, x_j) + \sum_i \alpha_i^*$$

The coefficient α_i^* define the maximal hyperplane in a high dimensional feature space, where the input data is mapped through $\sum_k \eta_k K_k(x_i, x_j)$. The gradient of the $J(\boldsymbol{\eta})$ in this

case is defined as

$$\frac{\delta J}{\delta \eta_k} = -\frac{1}{2}\sum_{i,j} \alpha_i^* \alpha_j^* t_i t_j K_k(x_i, x_j) \tag{3.4.5}$$

3.4.2 DAG based Classifier Design

Conventionally, the extension of SVM for the multi-class problem is done by decomposition based methods i.e. the problem is decomposed into set of a binary classification problems. Multi-class labelling is performed by a combinatorial use of the binary-classifiers. The two most preferred methodologies in this context are winner-takes-all using 1-Vs-rest binary classifiers and majority vote using 1-Vs-1 binary classifiers. However, classifier training with a large data in the case of 1-Vs-rest binary classifiers is computationally costly. In the case of multi-class MKL by binary 1-Vs-rest MKL classifiers, the computational cost multiplies due to the joint optimization procedure. The 1-Vs-1 methodology requires $N(N-1)/2$ binary SVMs trained for each pair of classes in N class problem. The labelling is done by applying the test point to all the binary classifiers and assigning the label from class set which gets maximum vote. Character/symbol recognition is described as a large multi-class problem where the application of $N(N-1)/2$ binary classifiers for final labelling is unacceptable in practice. The Decision DAG (Direct Acyclic Graph) framework proposed by Platt *et al.* [245] presents an efficient solution for combining the results of 1-Vs-1 binary SVMs for such large category problems. The framework arranges all the binary SVMs in a DAG architecture with $N(N-1)/2$ nodes in the graph (the graph for a 4-class problem with symbolic binary classifiers is shown in Figure 3.7). In the proposed framework, we arrange the set of binary MKLs in a DDAG framework. Figure 3.7

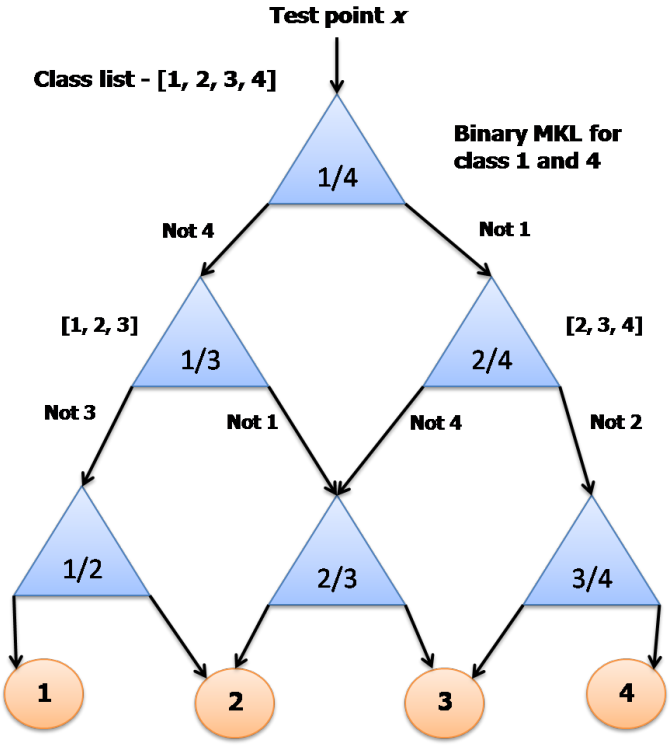


Figure 3.7: 4-class classification with binary MKL in DAG architecture

demonstrates the labelling process for a test point x in 4-class problem in DDAG framework. For a N class problem, the process evaluates $N-1$ nodes; therefore, significantly reducing the required number of kernel computations. The path followed by a DDAG for a test point labelling is called its evaluation path. The average kernel computation for complete test data is obtained by averaging over the count of unique support vectors over the evaluation path for all test points.

The MKL formulation presented in equation (3.4.3) is applicable for binary classification. Considering the conventional approach of decomposing the multi-class problem into set of binary problems, the possible extension of equation (3.4.3) is to define a global

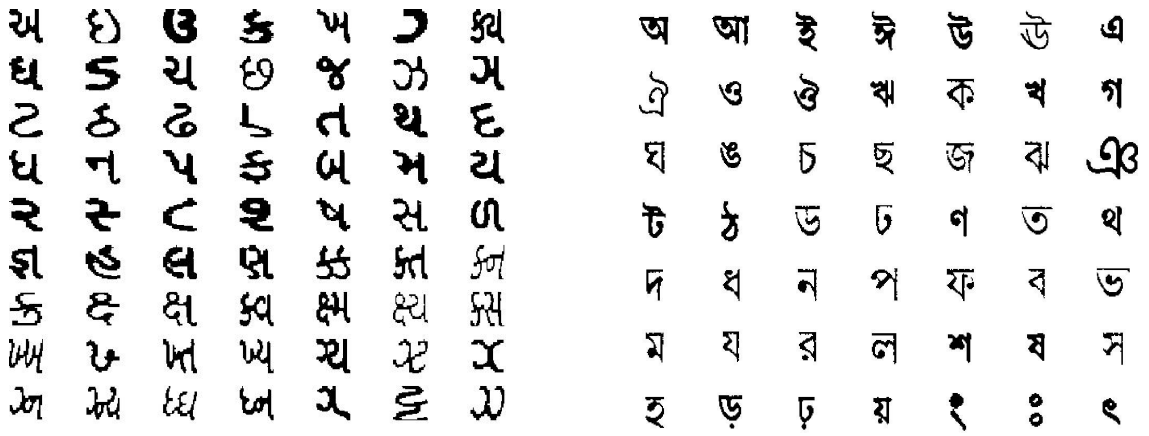
optimization problem $J(\boldsymbol{\eta})$ for joint optimization corresponding to all the binary classifiers. The objective function $J(\boldsymbol{\eta})$ is defined as $\sum_k^N J_k(\boldsymbol{\eta})$. Since the global objective function is a direct summation of individual $\{J_k : \forall k = 1, ..., N\}$, the gradient defined in equation (3.4.5) is easily extendible by the principle of linearity. However, the global optimization problem is not applicable in case of the DDAG architecture for multi-class MKL, as the test point labelling process does not include all binary classifiers. Therefore, the linearity assumption is not valid. Alternately, learning binary MKLs for all possible pairs of classes is another solution. This approach seems more intuitive as the kernel matrix is most informative when it is aligned with the target variable. In addition, selection of unique $\boldsymbol{\eta}$ for all the binary classifiers is not justifiable as the decision plane corresponding to a classifier is optimal with respect to its kernel space representation.

3.5 Experimental Evaluation and Discussion

The section presents the evaluation of proposed concepts for two applications: primitive/character recognition and symbol recognition. The experiments first evaluate the individual effectiveness of the different features. The subsequent experiments apply the proposed classification framework over different combination of features.

3.5.1 Character/Primitive Recognition

The first part of experiments are performed on Gujarati and Bengali character/primitive recognition. Figure 3.8 shows sample primitives from the experimental image collection.



(a) Gujarati script

(b) Bengali script

Figure 3.8: Sample character/primitive images

A common set of parameters has been applied for computing the feature representations corresponding to both the dataset. The shape descriptor, Fringe Map and HOG feature is used for primitive representation. The parameter description of the features is as follows:

* Fringe map for the example character/primitive images are computed after object bound detection and resizing to 32×32.

* A Shape descriptor is computed after object bound detection and resizing to 32×32 with logical grid S is selected of size 16×16. The histogram parameters m and n are selected as 35 and 36.

* A HOG feature is computed after object bound detection and normalization to 32×32 with each *cell* covering rectangular area of 8×8 pixels. The local histogram compu-

tation is done for 9 bins and the *block* level normalization is performed by 4 adjacent *cells* arranged as 2×2. Therefore, the HOG computation results in a vector of 144 elements.

The Gujarati example set used for experiments contains 5, 7 and 240 primitive categories form the lower, upper and middle zone respectively. The example image distribution corresponding to the three zones is as 457, 1307 and 13083 images respectively. The primitives are glyphs representing half or full forms of consonants, vowels or their combinations as discussed in section 3.2. The distribution shows that the majority of primitive classes originate from middle zone. Additionally, in general, the average support for each category is least for middle zone primitives which require major attention for feature extraction and classifier design. Therefore, the major objective of experiments is to improve the classification of middle zone primitives. For all the experiments, the feature sets have been used in the original form without any scaling or dimension reduction. The results presented in this section are the average of 5-fold cross validation. The SVM parameter tuning is performed by grid-based search and Euclidean distance is used for similarity measure in Nearest neighbour based classification. Initially, the discriminative power of feature sets is established by applying them for lower and upper zone primitives for classification. The results with KNN and SVM classifier are presented in table 3.1. The application of different features achieved classification accuracy between 95.89-99.32%. We accept the present results for these primitive categories and evaluate the middle zone primitive classification using individual features. The results in table 3.2 shows a similar order of accuracy using all the features.

Table 3.1: Gujarati characters: Classification of lower and upper zone primitives

	Lower zone primitives		Upper zone primitives	
	KNN (K=5)	SVM (DAG)	KNN (K=5)	SVM (DAG)
SD	95.89	97.79	96.21	98.13
FM	97.11	98.43	97.87	98.97
HOG	98.23	98.87	98.76	99.32

Table 3.2: Gujarati characters: Classification of middle zone primitives with KNN and SVM

	KNN (K=5)	SVM (DAG)
SD	93.36	95.57
FM	94.63	96.74
HOG	95.85	97.14

In the next step, the features are applied for recognition using the proposed MKL based classification. In this case, the base kernels for shape descriptor included linear, 2^{nd} order polynomial and 8 Gaussian kernels with variance ranging from $\{2^{-3}, ..., 2^{4}\}$. Similarly the base kernels for Fringe Map included 19 Gaussian kernels with variance ranging from $\{2^{0}, ..., 2^{9}\}$ and the base kernels for HOG included a set of linear and 16 Gaussian kernels with variance ranging from $\{10^{0}, ..., 10^{1.5}\}$. In all the experiments discussed here, the exponent of variances are scaled on linear scale with uniform step defined by the minimum and maximum and number of kernels. The results in table 3.3 shows improvement of 0.43-1.08% in accuracy over NN and SVM based performance. Also the HOG feature based classification is faster than other features.

For learning the combination of features, the base kernel set is formed by the union

Table 3.3: Gujarati characters: Classification of middle zone primitives with MKL

	MKL-DAG		MKL-1-Vs-1	
	Accuracy	Kernel comp.	Accuracy	Kernel comp.
SD	96.41	5205	96.37	8216
FM	97.13	5701	97.16	8823
HOG	97.61	3487	97.54	6365

of individual base kernels. In the case of combination of features using nearest neighbour classifier, the feature representation is defined by concatenating different features. Initially pairwise features are selected for MKL based classification. The results are presented in table 3.4. The combination of Shape Descriptor and Fringe Map increased the classification accuracy by 1.03% compared to the individual best (Refer table 3.3 and 3.4). The

Table 3.4: Gujarati characters: Classification of middle primitives by pairwise feature combination using MKL

	KNN	MKL-DAG		MKL-1-Vs-1	
		Accuracy	Kernel comp.	Accuracy	Kernel comp.
SD and FM	95.33	98.18	4315	98.21	7643
SD and HOG	95.77	97.86	3798	97.92	6871
FM and HOG	96.65	98.84	4273	98.74	7465

complementary nature of information represented by the feature set is efficiently combined by MKL. The L_1 norm constraint over $\boldsymbol{\eta}$ forces some η_k to zero; therefore, selecting only a few kernels for combination from the base kernel set. In the analysis of the kernel weight parameter $\boldsymbol{\eta}$ for the classifiers applied in the evaluation path of a test point, the average of the sum of η_k corresponding to each feature represents its contribution in

learning the optimal kernel space. The measure gives insight into the final classifier design as the estimate of information supplied by different features. In the present case, Fringe Map contributed by 67.8% weight following the discussed procedure. The MKL based combination of the shape descriptor and HOG increase the accuracy by a small margin of 0.17% over individual best. The reason for the small improvement is the exploitation of similar symbol characteristics for feature computation as both the features fundamentally represent shape. The HOG feature is the dominating partner contributing 96.5% weights in final combination. The combination of HOG with Fringe Map improved the classification accuracy by a reasonable margin of 1.23% over the individual best. The result establishes the claim of efficient utilization of complementary information inbuilt in different features for performance improvement. The HOG feature again is the dominating partner contributing 87.3% weights though the dominance is lower than the observation in combination of HOG and shape descriptor. However the combination required a marginally increased number of kernel computations.

Finally, combination of Shape Descriptor, Fringe Map and HOG feature is learned for recognition. The result in table 3.5 shows that combination improved classification accuracy by 0.19% with marginal increase in average kernel computations (2.41%) than pair-wise best results. The observation of $\boldsymbol{\eta}$ showed that HOG and Fringe Map contributed 87.5% and 9.8% kernel weights. Additionally, we observe that the distribution of kernel weights in different feature combinations have followed the trend of $\eta_{SD} < \eta_{FM} < \eta_{HOG}$, which is in accordance with their performance with base classifiers (table 3.2). The experimental results show that multiple features based primitive representation combined with

Table 3.5: Gujarati characters: Combination of Shape Descriptor, Fringe Map and HOG

KNN	MKL-DAG		MKL-1-Vs-1	
	Accuracy	Kernel comp.	Accuracy	Kernel comp.
97.09	99.03	4376	98.98	7596

MKL based classification provides a robust classification framework for OCR applications. It must be observed that the features discussed above contain reasonable amount of overlapping information, however the MKL utilizes the complementary information from different features to learn the resulting feature space. The mean accuracies presented in table 3.3, 3.4, and 3.5 based on MKL classification have well defined significance. Nevertheless, the observed difference between average means may have been generated by chance. In order to conclude the absolute mean difference, we also need to consider the within group variability. In particular, if within group variation is significantly smaller than inter-group variation, we conclude that observation has a real effect. For such statistical analysis, a hypothesis test would further establish presented classification results. Here the null hypothesis says that mean accuracies for different groups are same. Cross-validation based procedure generates set of accuracies (or selected performance measure) by considering each fold of the dataset as testing set. Here, we perform one-way ANOVA (Analysis of Variance) to test the difference between groups of accuracies obtained for different feature combinations. The procedure produces one-way analysis of variance for classification accuracy with respect to used feature sets. Table 3.6 presents the one-way analysis of variance of cross-validation accuracies obtained by different feature and their

combinations. The p-values obtained for both classification configurations i.e., DAG and 1-Vs-1 are significantly below significance level of 0.05 which establish real difference between the presented mean accuracies.

Table 3.6: One-way ANOVA Table: Input Groups - {MKL based cross-validation accuracies using MSD and FM, HOG and FM, HOG and MSD, HOG and MSD and FM}, SV - Source of Variation, SS - Sum of Squares, df - Degree of Freedom, EV - Empirical Variance

Feature combinations in MKL-DAG			
SV	SS	df	EV
Between Groups	16.1629	3	5.3876
Within Groups	6.3515	36	0.1764
Total	22.5144	39	
Test Statistic (F)	30.5367		
p-value	0.0000		
Feature combinations in MKL 1-Vs-1			
SV	SS	df	EV
Between Groups	19.1293	3	6.3764
Within Groups	6.6982	36	0.1861
Total	25.8274	39	
Test Statistic (F)	34.2706		
p-value	0.0000		

The Bengali character recognition experiment is performed on image collection consists of 17000 example images of 49 categories. The examples represent isolated Bengali alphabets (vowels and consonants) which appear in the middle zone of the word object. Therefore, images corresponding to *Chandra Bindoo* are not considered for dataset compilation. The experiments estimate the improvement in classification accuracy by the

application of multiple features by MKL. Similar experimental methodology is adopted as discussed for Gujarati script. First we evaluate the character classification accuracy using KNN and SVM as base classifiers. The SVM parameter tuning is performed by grid based search following 5-fold cross validation. The results are presented in the table 3.7.

Table 3.7: Bengali characters: Classification using KNN and SVM

	KNN (K=5)	SVM (DAG)
SD	94.63	95.56
FM	95.14	96.07
HOG	98.28	98.63

In following step, the proposed MKL framework is applied for classification. The base kernel selection for each feature is as following: 2^{nd} order polynomial and 17 Gaussian kernels with variance ranging from $\{2^{-1}, ..., 2^{7}\}$ for shape descriptor, linear and 17 Gaussian kernels with variance ranging from $\{2^{2}, ..., 2^{8}\}$ for Fringe Map, and set of linear and 16 Gaussian kernels with variance ranging from $\{10^{0}, ..., 10^{1.5}\}$ for HOG. The results with individual features are shown in the table 3.8. The results using HOG feature is distinctly better using base classifier as well as MKL based classifier. However, we observe that accuracy improvement is less than other features (0.44%, than 1.20% and 0.96% in case of Shape Descriptor and Fringe Map in DAG architecture). In this case, Fringe Map based classifier requires least kernel computations while HOG based classifier performs 0.06% more computations in comparison. The MKL based character classification using HOG feature has shown reasonably acceptable performance ($\geq 99.07\%$). However to investigate further improvement in accuracy, we first learn combination of Shape Descriptor

Table 3.8: Bengali characters: Classification using MKL

	MKL-DAG		MKL-1-Vs-1	
	Accuracy	Kernel comp.	Accuracy	Kernel comp.
SD	96.47	2303	96.38	4094
FM	96.88	1298	97.02	2166
HOG	99.07	1378	99.11	2432
Combination of Shape Descriptor and Fringe Map				
KNN	MKL-DAG		MKL-1-Vs-1	
	Accuracy	Kernel comp.	Accuracy	Kernel comp.
96.63	98.13	1483	98.08	2524
Combination of Shape Descriptor, Fringe Map and HOG				
KNN	MKL-DAG		MKL-1-Vs-1	
	Accuracy	Kernel comp.	Accuracy	Kernel comp.
98.16	99.48	1371	99.41	2237

and Fringe Map using MKL. In this case, the base kernel set is formed by union of individual base kernels. In case of nearest neighbour, the feature representation is defined by concatenating different features. The results in table 3.8 shows that optimal combination of both the features has increased the classification accuracy 1.25% in comparison with individually best. Nevertheless the process requires 14.25% more kernel computations. The observation of kernel weight parameter $\boldsymbol{\eta}$ showed that, on average 73.53% weight belonged to Fringe Map. The combination of HOG with other two features improved average classification accuracy by 0.41% at computational cost of the similar order of when HOG is applied independently. The examination of $\boldsymbol{\eta}$ showed that for the final kernel used for classification, 89.11% kernels belonged to HOG and 6.17% belonged to Fringe Map. The result is obvious as the classification accuracy with HOG feature has been significantly

high using SVM and MKL in comparison with other features. However by combination of features using presented MKL framework, we are able to improve the character classification accuracy by 0.85%. Results presented in table 3.8 clearly show the improvement in classification accuracy using combination of features. The MKL classification based cross-validation accuracies obtained for different features and their combinations shown in table 3.8 have been further validated using one way analysis of variance based significant test. Table 3.9 shows the comparative measures where the null hypothesis is rejected with significant confidence as shown by the obtained p-values.

Table 3.9: One-way ANOVA Table: Input Groups - {MKL based cross-validation accuracies using SD, FM, HOG, SD and FM, SD and FM and HOG}, SV - Source of Variation, SS - Sum of Squares, df - Degree of Freedom, EV - Empirical Variance

Feature combinations in MKL-DAG			
SV	SS	df	EV
Between Groups	33.9173	4	8.4793
Within Groups	3.0548	20	0.1527
Total	36.9721	24	
Test Statistic (F)	55.5143		
p-value	0.0000		
Feature combinations in MKL 1-Vs-1			
SV	SS	df	EV
Between Groups	37.8362	4	9.4590
Within Groups	2.6412	20	0.1321
Total	40.4774	24	
Test Statistic (F)	71.6260		
p-value	0.0000		

The above experiments demonstrate significant improvement in baseline (NN and

SVM based results). The classification accuracy achieved for Gujarati character recognition is significantly better in comparison with the results presented in [73]. In addition, the framework has shown robust performance while considering more number of primitive categories. Bengali character recognition is much researched topic. Our framework has shown significant improvement over the baseline results. Additionally, the comparison of our results with [239, 196, 109] show comparable or improved accuracy. Our framework presents an efficient classification approach for character recognition by efficient combination of structural and shape based feature sets. Additionally, the classification framework is much faster than conventional framework which is an essential requirement for OCR based applications.

3.5.2 Symbol Recognition

The experiments for the symbol recognition is performed on the **MPEG-7 CE Shape-1 Part-B** dataset available at [302]. The collection consists of 70 symbol categories having 20 examples each. The examples exhibit significant variations covering translation, rotation, scaling and non-rigid deformations. First, different features discussed in the section 3.3 are applied for symbol representation in NN and SVM based classification. All the related experiments have been performed using cross-validation over 10-folds. A NN based classification is performed with Euclidean distance as the similarity measure and 3 nearest neighbours are considered for majority voting. The feature extraction details for different representations are listed below.

* The Fringe Map for the example images are computed after symbol bound detection and normalization to 64×64.

* The Shape descriptor is computed after symbol bound detection and normalization to 128×128 with logical grid S of size 32×32. The histogram parameters m and n are selected as 40 and 36.

* The HOG feature is computed after symbol bound detection and normalization to 128×128 with each *cell* covering rectangular area of 32×32 pixels. The local histogram computation is done for 24 bins and the *block* level normalization is performed by 4 adjacent *cells* arranged as 2×2.

Initial classification results presented in table 3.10 establish the effectiveness of HOG feature in addressing different shape distortions existing in symbol images. The localized approach for HOG computation incorporates strong invariance to minor degradations and distortions affecting parts of symbol shape.

Table 3.10: Symbol classification using individual features

	KNN	SVM (DAG)
FM	76.68	78.92
HOG	83.71	84.32
SD	73.45	74.87

In the following step, the modified shape descriptor is applied for symbol image representation. The descriptor is computed after normalizing the bounded symbol image to 128×128. The logical grids S and S_{centre} are of size 32×32. The histogram parameters

m and n are selected as 40 and 36. The classification results are shown in table 3.11. The descriptor computation after smoothing the sampled boundary points improves its robustness. In general, inner contours are not very common in symbols images compared to the character images. In this case, the descriptor primarily represents global shape information. Nevertheless, the original shape descriptor is more sensitive to distortions, although the margin is reduced after smoothing the descriptor point set. The results establish the effectiveness of the modified shape descriptor where the NN based results are comparable with the recent result presented in [84]. Additionally, the SVM based classification achieved substantial improvement in the accuracy.

Table 3.11: Symbol classification with modified shape descriptor

	KNN	SVM (DAG)
MSD	80.79	82.08

In the following experiment, multiple features are applied for symbol classification. The proposed framework learns an optimal combination of different representations for the recognition. First, the pair-wise combination of the modified shape descriptor, fringe map and HOG is used. Next, all the features are combined for recognition. The base kernel selection for each feature set is as following: linear and 13 Gaussian kernels with variance ranging from $\{2^{-1}, ..., 2^{6}\}$ for modified shape descriptor, linear and 15 Gaussian kernels with variance ranging from $\{2^{1}, ..., 2^{7}\}$ for Fringe Map, and set of linear and 11 Gaussian kernels with variance ranging from $\{2^{-1}, ..., 2^{3}\}$ for HOG. The results are presented in table 3.12. The pairwise combinations of modified shape descriptor with Fringe Map,

Table 3.12: Symbol classification using combination of features by MKL

	MKL-DAG		MKL-1-Vs-1	
	Accuracy	Kernel comp.	Accuracy	Kernel comp.
MSD and FM	84.36	732	84.47	1643
HOG and FM	85.64	865	85.76	1895
HOG and MSD	84.41	774	84.38	1757
HOG, MSD and FM	85.48	844	85.54	1901
1-Vs-1 Error-correcting Output Codes Scheme with Gentle Adaboost and Circular Blurred Shape Model as reported in [84]				80.36

and HOG with Fringe Map show significant improvement over individual best results. The combined application of HOG and modified shape descriptor does not improve the performance typically because of the overlapping nature of information. The combination of HOG with Fringe Map achieved best result with 88.63% kernels contributed by HOG. Next, the combination of three features for recognition achieved comparable result to the combination of HOG with Fringe Map. The analysis of kernel weight parameter η shows that HOG descriptor is the dominant contributor supplying 75.94% kernels with modified shape descriptor contributing 17.63% kernels. The results in table 3.12 show 5.28% improvement in the classification accuracy over the results obtained in [84]. Additionally, the features proposed in this work are much easier to compute, and classifier training and prediction process is much simpler and straight forward. Again the DAG based formulation is much faster for recognition. We have demonstrated that recognition performance is significantly improved by principled combination of simple features. HOG descriptor characterizes object shape information by orientation histograms computed in

local neighbourhood. The modified shape descriptor considers global as well as local shape information for description. The fringe map is continuous descriptor and represents the structure of the object by extracting the distance information between pixels. These complementary informations are efficiently combined by our MKL based classification framework which is established by the experimental analysis. Table 3.13 presents the one-way analysis of variance of cross-validation accuracies obtained by different feature combinations. For both classification configuration i.e., DAG and 1-Vs-1, we observe that p-values is significantly below significance level of 0.05. With such high value of F-statistics, the results strongly indicate the real difference between mean classification accuracies for different feature combinations.

3.6 Conclusions

The chapter presented novel classification framework for binary pattern recognition. The applicability of the framework is demonstrated for character/primitive labelling for Gujarati and Bengali character recognition. The generalization of the framework is shown for symbol recognition. We presented DAG based architecture of MKL for a large-class categorization problem for addressing the requirement of fast recognition. The experimental evaluation showed that the framework presents a robust method for binary pattern recognition by optimally combining multiple features. Additionally, a novel feature i.e. shape descriptor is proposed for binary pattern representation. The comparative analysis of character and symbol classification results show that a boundary information based

Table 3.13: One-way ANOVA Table: Input Groups - {MKL based cross-validation accuracies using MSD and FM, HOG and FM, HOG and MSD, HOG abd MSD and FM}, SV - Source of Variation, SS - Sum of Squares, df - Degree of Freedom, EV - Empirical Variance

Feature combinations in MKL-DAG			
SV	SS	df	EV
Between Groups	16.1629	3	5.3876
Within Groups	6.3515	36	0.1764
Total	22.5144	39	
Test Statistic (F)	30.5367		
p-value	0.0000		
Feature combinations in MKL 1-Vs-1			
SV	SS	df	EV
Between Groups	19.1293	3	6.3764
Within Groups	6.6982	36	0.1861
Total	25.8274	39	
Test Statistic (F)	34.2706		
p-value	0.0000		

descriptor provides robust feature representation option for binary patterns. Here, we note that, the modification in the shape descriptor for addressing elastic deformations in symbols with other distortions demonstrated significant improvement over the recent symbol recognition results presented in [84]. The recognition performance using the shape descriptor representation motivated for its application for the development of word image based document indexing framework presented in the chapter 4. Here, we have proposed MKL for large class problem using DAG architecture. However, the framework does not exploit the complete knowledge of the data because of the 1-Vs-1 nature of learning. An

interesting extension as part of future work could be the exploration of novel mathematical formulation for MKL for such large-scale applications having large number of symbol categories.

Chapter 4

Word based Document Image Indexing and Retrieval

4.1 Introduction

The digitization of documents across the world has created a large collection of document images. Indexing of these document images poses a challenging problem. In the traditional document image indexing systems, optical character recognition (OCR) is applied to convert the document image to an electronic text representation. The recognized characters/words are further used to build an indexing scheme for documents. The precondition for the approach is the availability of robust optical character recognizer for the script. However, for old and degraded documents, and for the scripts for which reliable OCRs are not available, this approach can not be followed. In such situations, we need to use image based indexing and retrieval schemes for the document collection. The indexing

scheme then exploits the image properties of the document content. The word image based document indexing and retrieval provides a practical solution for indexing the non-OCRed document collections. The textual contents in the document, i.e., word images are used for generating indices. The primary challenges involved in defining a word image based document indexing framework are (i) formulating a unique feature based representation for word images, and (ii) developing computationally inexpensive method for indexing large collection of document images. The work presented in this chapter addresses both the issues related to word based document indexing. Following are the major contributions in this chapter.

- The word images extracted from the document images are example binary patterns. In the section 3.3.3, we have presented a shape descriptor feature for binary patterns. In this chapter, we explore the applicability of shape descriptors for word image representation and present its extension for word recognition and retrieval.

- A novel word based indexing and retrieval framework for document image collection is developed by applying an enhanced distance based hashing (DBH). The distance based hash functions are binary mapping functions defined over a cosine law based line projection. In DBH, the hashing functions are learned through word images belonging to a document collection. The following are distinct features of our work:

  - In conventional DBH, the learning of hashing functions is performed over a set of randomly selected word objects. We present a computation of hashing functions based on precomputed cluster centres of the training data such that

the collision probability (probability of hashing of similar objects to the same location) is increased.

  - We formulate a hierarchical DBH scheme for document indexing which reduces the retrieval search complexity by maintaining hierarchical hash tables over base hashing. The shape descriptor based representation is applied for word images in indexing and retrieval framework.
  - The novel concept of multi-probe hashing is extended for binary mapping functions. We demonstrate the applicability of the proposed Multi-probe hashing using DBH functions.

- Additionally, a modified document image indexing and retrieval framework is proposed which uses a string based representation for word images. The indexing framework follows the conventional distance based hashing where edit distance based similarity is used for generating the word indices.

The experimental evaluation of presented concepts have been shown on document collections belonging to Devanagari, Bengali and English scripts. First, we review the existing word image representation in following section. Subsequently, a DBH based document indexing scheme is introduced.

4.1.1 Analysis of Feature Representations for Word Images

The available research in the area of word image representation have discussed character-like coding and shape based representation as two primary strategies. The character-like

coding representation is generated by extracting some objects from the word image through classical or morphological segmentation. Each object is assigned a code based on shape similarity to labelled data. The codes are concatenated in the respective order of objects to generate the word image code. In recent work, Shijian *et al.* [193] have proposed a word shape coding scheme which considers a word as a single component that does not require character segmentation. It defines a word image by a set of topological character shape features including character ascenders/descenders, character holes and reservoirs. Simone *et al.* [205] have defined collection specific character prototypes from the character objects. The set of character prototypes are further used for word image representation. Nakayama [223] have defined word shape tokens for printed word images by a sequence of character shape codes. The character shape codes are defined by the set of graphical features. In [20], a word shape code is generated based on standard features as ascenders, descenders, character holes, deep eastward and westward concavity and horizontal-line intersections. The character like coding schemes for word images are very much script specific and generalization of these schemes for different scripts is a difficult task. In addition, the detection of various word image features (topological and morphological) is very sensitive to noise and document image quality. Howe et al. [139] proposed application of pixel points in pyramid structure generated for the normalized handwritten word image. Rath and Manmatha applied a set of profile based features in combination with dynamic time warping based matching for handwritten document retrieval [256, 255]. Konidaris et al. [156] presented combination is zone based profile based features for word image representation in historical printed documents. Zone based charactersitics is represented

by density of character pixels in each zone, profile based characteristics is represented by area formed from the projections of the word image. Kesidis1 and Gatos further used theses set of features for threshold based word spotting application described in [151].

The shape based schemes utilize the appearance of the word as an entity, using its features like outer boundary, horizontal and vertical lines, inner circles and sharp turns for representations. The earliest work on word spotting presented in [52] has used word contours in addition with auto correlation coefficients as features. Madhavanath *et al.* [195] have followed a holistic approach for handwritten word images representation for addressing the character segmentation in handwritten document recognition. In [21], envelope curve based signatures of the handwritten word images are used for signature verification. The envelopes are derived as a sequence of external points with respect to the principal axis of the signature. Along similar lines, contours of the handwritten word image have been used for representation in [3]. Recently, Wshah et al. [336] presented novel methodolgy for word spotting in offline handwritten documents by simulating all the keywords by combination of HMMs learned on trained characters. In [160], discrete Fourier transform coeffcients based word image representation is applied for document indexing and retrieval. The earliest method for text based handwritten document retrieval presented novel usage of transform coefficients of word shape profile for representation [257]. Meshesha and Jawahar [213] presented extensive evaluation of set of local and global features including profile, moments, and tranform based features for word based document retrieval. Bhardwaj et al. [27] demonstrated the application of geometric moments for sanskrit word image representation. Gatos and Pratikakis [110] developed a set of heuristic based

feature vectors by applying rotation and scaling operation. Howe and Manmatha [138] used hitogram of gradients for handwritten character detection. The character sequence is further learned by ensemble of hidden Markov models for word recognition. Praveen et al. [246] proposed word image representation as bag-of-character n-grams which are represented by profile based features in visual-feature space. Recent work in this direction have explored application of bag-of-words based word image representation for different Indian scripts using local gradient features [157, 275].

The recent development of shape context descriptor has performed excellently for various recognition problems [25]. However, the computational complexity of similarity search using rich set of local descriptor represented by shape contexts is very high. In [190], the skeleton of a word image has been used to represent word shape. The word shape signature is formulated by computing the shape context over the points sampled over the skeleton of word image. However the skeletons are highly noise sensitive. In general, shape based features define geometric relations between the set of points over the word object. The feature represents the visual characteristics of word shape by these geometric relations. Therefore, the exploitation of these features provides more intuitive and flexible approach for defining script independent word image representation. However, in the case of degraded and low quality document images, techniques such as contour detection and skeletonisation for shape information extraction have serious limitations. In contrast, the proposed feature extraction scheme has the advantage of grid based approach for shape information extraction which does not require edge, contour or skeleton detection. The approach implicitly handles general scenarios and can distinguish words having similar

outer envelopes but different inner contours. Additionally, the feature represents word objects as constant dimensional vectors which give flexibility to use direct comparison methods for similarity matching.

4.2 Overview of the Document Indexing Framework

In this section, an overview of the document image retrieval system is presented. The document indexing scheme follows a conventional hashing based indexing framework (Figure 4.1). Here we apply distance based hashing functions for generating the indices.

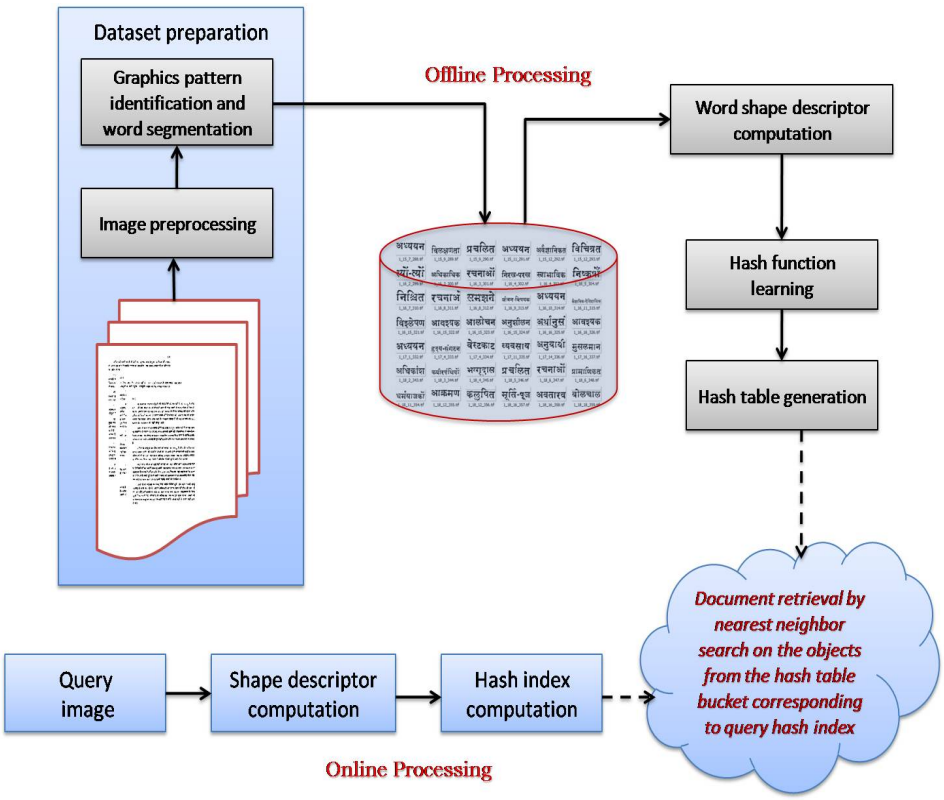


Figure 4.1: Document indexing framework

The initial steps of off-line processing include document preprocessing (deskewing and binarization). The segmentation routine extracts word images from the document image collection. The meaningful word images for generating the document indices are selected using word length in terms of pixels as a thresholding criterion, e.g, most of the English words having less than four characters are not required for indexing. These words in addition with stop words, punctuation and typographical marks are filtered out by applying conventional thresholds (aspect-ratio, word length). The feature representation for word images is defined by exploiting the distinct image attributes with supported distance measure for establishing the similarity of two words. The next step of off-line process includes the generation of distance based hash functions. Each hashing function is defined for a pair of data points defined as pivots. Instead of a conventional approach, we consider precomputed cluster centres obtained by clustering the word images, as pivot objects. These pivot objects are used to generate a family of hashing functions $\mathbb{H}$. The hashing function generation complexity is $O(\beta^2 N^2)$. Here N is the word image count, and β is the ratio of number of clusters to N. For L hash tables we generate g_i for $i = 1, ..., L$ by random selection of k functions from $\mathbb{H}$. Using each function g_i, the indices for word images belonging to complete collection is generated and the procedure is repeated for all the hash tables. Each separable set obtained after the application of g_i over the set forms a bucket. In this respect a bucket remains a metric region and organizes all word images from the metric domain falling into it. Each hash table requires $O(kN)$ distance calculations (each calculation signifies one projection operation). The storage architecture of the hash table is based on 2-dimensional array of buckets used for storing data points. This requires

$O(kLN)$ space to record the k-bit coordinate values corresponding to N word images for L hash tables.

The on-line process includes retrieval by performing the similarity search based on query. This includes a word image generation from the query text string. The next step computes shape descriptor for the query image. The query is indexed to all the hash tables using hashing function g_i. The data points from the query bucket (bucket to which query is hashed) in all the hash tables are collected. The word images similar to query word are retrieved by performing similarity search over the group of collected data points. Theoretically the DBH does not guarantee sub linear search complexity. Nevertheless, in practice, it achieves good match with reduced search complexity which is confirmed by the experimental results.

4.3 Shape based Feature Representation for Word Images

The following discussion describes the shape based feature extraction and representation scheme applied to word images.

4.3.1 Extension of Shape Descriptor for Word Image Representation

In the case of word objects, the character sequence information in the shape descriptor presented in section 3.3.3 suffers because of the global nature of h_{sum}. The empirical evaluations have shown that character sequence information can be incorporated by splitting the word image into a constant number of partitions. However segmenting a word

image into predetermined number of partitions is a challenging problem. In the context of Indian scripts, the problem is further compounded because of the modifiers above and below the characters in word formation (Figure 4.2 shows the use of modifiers for word formation). These problems make the character segmentation process highly error prone. Therefore, to make the shape descriptor invariant to symbol/character segments, we split the word image into partitions of equal width (e.g. the word image is partitioned into four parts in figure 4.3). For each partition, the h_{sum} is computed independently with points

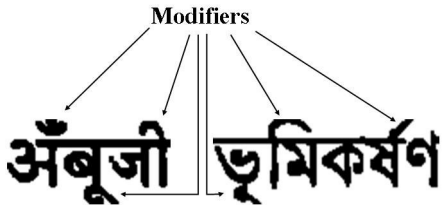


Figure 4.2: Modifiers on the word image

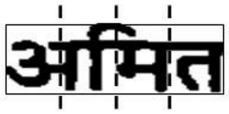


Figure 4.3: Partitions on the word image

corresponding to the partition. The final histogram h_{final} is obtained by concatenating normalized h_{sum} corresponding to all the partitions following their sequence. If there are num_parts partitions, the dimension of the final histogram h_{final} is $m \times n \times num\_parts$. Such an arrangement helps in preserving the sequence information in a word image, that otherwise suffers in global distribution. In addition, the deformations into partitions of a word image will not affect the pdh corresponding to other partitions. The isolated noisy bins from the histogram are filtered out by applying adaptive filtering. Small distortions in the word shape are handled by applying smoothing over an appropriate neighbourhood over the final histogram.

The selection of the number of partitions is based on heuristics by considering the

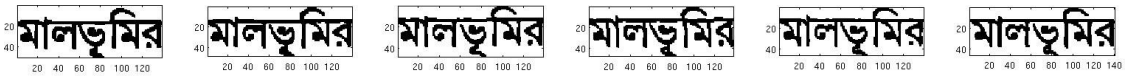
average number of characters in word formation as guidance, e.g., most of the Devanagari words are formed by combination of 3 to 5 characters; therefore, we can select 1×3, 1×4 or 1×5 partitions for splitting the word image. Every unique word is represented distinctly by a combination of global and local shape features. The global shape features define the overall shape of the histogram. The local shape features like internal contours, sharp curves and broken or faded characters contribute into the smoothness of the histogram. The feature representation for the word image is obtained by an application of Fourier transform over h_{final}. The shape descriptor computation steps are summarized as:

- Computation of h_{sum} for all partitions w.r.t. points in the corresponding partition
- $h_{final} = \{h_1 : h_2 : ... : h_{num\_parts}\}$, h_i represents the normalized h_{sum} for i^{th} partition
- The shape descriptor $F(P)$ is defined by the magnitude of the Fourier Transform of h_{final}

Since the final histogram is a sequence of independent histograms, the partially matching results can also be retrieved. The image sequence in figure 4.4 shows an example of retrieval based on the shape descriptor computation following without, and with 1×4 partitioning of word image.

In the case of similar words having different font properties, variations in the outer envelope of word image is not significant leaving the low frequency components of shape descriptor unchanged. Table 4.1 represents the average distance between five word images using the proposed shape descriptor (Figure 4.5). The descriptor parameters are $\{m =$

(a) Feature computation without partitioning



(b) Feature computation with 1×4 Partition

Figure 4.4: Nearest neighbour based retrieval: First image is the query and remaining images are ranked by Euclidean distance as similarity measures

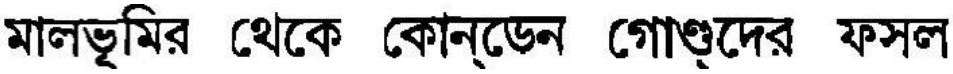


Figure 4.5: Sample images considered for computing the distance matrix

$40, n = 40$, 1×4 Partition}. The number of examples in each group of words are {15, 3, 30, 8, 10}.

Table 4.1: Distance matrix for the words shown in figure 4.5

Image	1	2	3	4	5
1	**0.20**	1.05	0.54	0.56	0.82
2	1.05	**0.16**	0.76	0.69	0.65
3	0.54	0.76	**0.16**	0.27	0.78
4	0.56	0.69	0.27	**0.16**	0.68
5	0.82	0.65	0.78	0.68	**0.23**

The discriminative property of the feature is established by observing the diagonal of the distance matrix. The distance distribution shows a close match between the third and fourth word in the sequence because of the approximately similar outer boundary.

Nevertheless the inner contour information in both the words represent them distinctly. The rotation invariance in the shape context can be incorporated in two ways. The first method considers the tangent vector at the point as reference axis for shape context computation. In the second method, the reference axis can be aligned with the principal axis of the shape to incorporate rotation invariance. The computation process of shape context makes it robust under small geometrical distortions and presence of outliers. The proposed feature represents the object image by constant length vector that can be applied in different applications. In addition, the feature gives the freedom of application of various vector based methods for performing similarity search.

The initial results on Bengali, Gujarati and Devanagari script characters and words establish the effectiveness of the shape descriptor. The shape based nature of the representation extends its applicability to different scripts in general. In the case of scripts having complex characters and modifiers e.g., the majority of South Indian scripts, a dense sampling of descriptor points is required. South Indian scripts exhibit a complex formation of word images having curly characters with no concept of horizontal line at the top. In this case, the histogram parameter selection requires careful analysis so that discriminative attributes of word shapes are discovered. Additionally, the South Indian words in general have more variation in terms of numbers of characters. Therefore, the selection of partitions should be such that deformations in word shapes are efficiently addressed at local level. Figure 4.6 shows the Telugu script retrieval with word image representation defined by proposed shape descriptor. The sampled document images are collection of old story book pages scanned at 300dpi [15]. The segmented word image collection consisted of

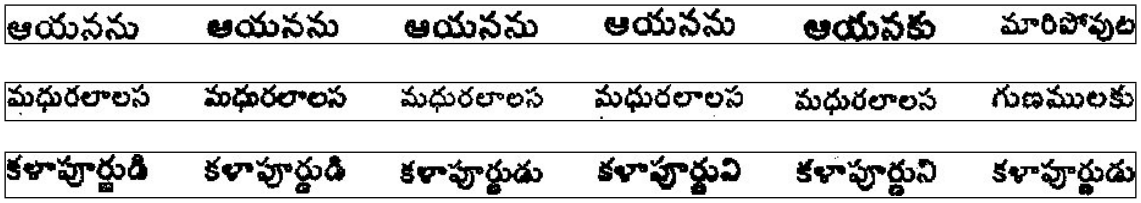


Figure 4.6: Sample Telugu script word retrieval: First image is query and remaining images are ranked on the Euclidean distance based similarity

7800 images. The descriptor is computed with $\{m = 38, n = 30, 1\times6$ Partition$\}$ with the logical grid placed at the interval of 4 pixels. The Euclidean distance is used for similarity measurement. Experimental results establish that proposed descriptor presents an efficient option for feature level representation of word images of different scripts.

4.4 Distance based Hashing for Indexing

Index space formed with shape descriptor is expected to be high dimensional. The inherent semantic structure of the object space can be explored by projecting the data onto a lower dimensional space. We use the concept of hashing for defining the lower dimensional representation of data. It is an efficient method to retrieve ϵ-approximate nearest neighbours, whose distance to query is atmost some factor $c = 1 + \epsilon \geq 1$ larger than the distance from the query to actual nearest neighbour.

A brief review of locality sensitive hashing is presented in Appendix A. A short introduction to distance based hashing is presented in order. The distance based hashing function is defined over a pair of pivot objects [311]. We propose clustering based method

for the selection of pivot objects for Distance based hashing. Subsequently, the locality sensitivity property of DBH is analysed using Euclidean distance as similarity measure. Finally, we introduce the novel concept of hierarchical Distance based hashing.

4.4.1 Distance based Hashing

LSH defines an indexing scheme: hashing data points using k hashing functions, and increasing the success probability of similarity search by generating multiple hash tables. Following the idea, Vassilis *et al.* presented the concept of Distance based hashing in [311]. Fundamentally, the idea comes from the FastMap embedding method [89]. The DBH is an algorithm to map objects to points in k-dimensional space such that the inter object distances are preserved. The distances represent the dissimilarity between objects. The DBH assumes that the objects are basically points in hypothetical Euclidean space with a defined distance measure. The heart of DBH is the projection of objects onto a carefully selected line while maintaining their distances. The line projection function defines one such mapping [89]. For two objects (x_1, x_2) in space $(\mathcal{X}, \mathcal{D})$ having objects represented as points of unknown dimension, the line projection $F^{x_1,x_2} : \mathcal{X} \rightarrow \mathfrak{L}$ for object x is defined as

$$F^{x_1,x_2}(x) = \frac{\mathcal{D}(x_1, x)^2 - \mathcal{D}(x_2, x)^2 + \mathcal{D}(x_1, x_2)^2}{2\mathcal{D}(x_1, x_2)} \tag{4.4.1}$$

$\mathfrak{L}$ defines the line connecting points (x_1, x_2). The only requirement for the function in equation (4.4.1) is the availability of distance $\mathcal{D}$; therefore the mapping F is also applicable for arbitrary spaces. The extension of mapping the objects to k dimensional

space is performed by projection using k-mapping functions. For every mapping function, two pivot objects (x_1, x_2) are chosen, a line is drawn between them that serves as coordinate axis, and the coordinate value along this axis for each object is determined by equation (4.4.1). In the case of $\mathcal{X}$ being a general non-Euclidean space, $F^{x_1,x_2}(x)$ is geometrically uninterpretable. However, if $\mathcal{D}$ is available for $\mathcal{X}$, F^{x_1,x_2} can be defined which provides a simple way to project x on the line defined by (x_1, x_2). The mapping F is independent of the dimensionality of the object representation as the inter object distance is the only requirement. Equation (4.4.1) defines a rich family of projection functions. For a collection of N objects in $\mathcal{X}$, $N(N-1)/2$ unique functions can be defined by applying equation (4.4.1) to each pair of objects. In practice, it is always convenient to have a hashing function that maps objects to $\{0, 1\}$. The functions defined using equation (4.4.1) are real valued, whereas we desire binary hashing functions. The binary hashing functions can be obtained from F^{x_1,x_2} using thresholds $t_1, t_2 \in R$ as:

$$F^{x_1,x_2}_{t_1,t_2}(x) = \begin{cases} 1 & \text{if } \; F^{x_1,x_2}(x) \in [t_1, t_2] \\ 0 & \text{otherwise} \end{cases} \tag{4.4.2}$$

The mapping defined in equation (4.4.2) can also be extended for step-wise projection. In practice, the selection of $[t_1, t_2]$ should be such that $F^{x_1,x_2}_{t_1,t_2}(x)$ maps approximately half the data points in $\mathcal{X}$ to 0 and the remaining to 1, i.e. F generates balanced hash tables. Formally for each pair $(x_1, x_2) \in \mathcal{X}$, the set $V(x_1, x_2)$ of intervals $[t_1, t_2]$ is defined such that $F^{x_1,x_2}_{t_1,t_2}(x)$ splits the hash space in half as

$$V(x_1, x_2) = [t_1, t_2] | \mathrm{Pr}_{x \in \mathcal{X}}(F^{x_1,x_2}_{t_1,t_2}(x) = 0) = 0.5 \tag{4.4.3}$$

With the threshold parameter family $V(x_1, x_2)$, we define the family $\mathbb{H}_{\mathbb{DBH}}$ for an arbitrary

space $(\mathcal{X}, \mathcal{D})$ as

$$\mathbb{H}_{\mathbb{DBH}} = F_{t_1,t_2}^{x_1,x_2}(x) | x_1, x_2 \in \mathcal{X}, [t_1, t_2] \in V(x_1, x_2) \tag{4.4.4}$$

An indexing scheme is formulated by defining g by randomly selecting k functions from $\mathbb{H}_{\mathbb{DBH}}$, and using it to generate a hash table for word objects. The retrieval success rate is increased by generating L hash tables. Retrieval is performed by hashing the query and collecting objects from all tables for similarity search. The implication of hashing parameters (L, k) in DBH is the same as for the LSH based indexing.

4.4.2 Pivot Object Selection

The selection of pivot objects (x_1, x_2) is an important issue for function F (Equation 4.4.1). In this section, we present the proposed scheme for pivot object selection. Ideally the pivot objects should be such that, the projection values are well separated on the connecting line. The underlying distance information between objects can be extracted more efficiently by greater spread between the pivot objects. The determination of the farthest pair of objects among a given set of N objects needs $O(N^2)$ distance computations. To reduce the computation cost, Faloutsos and Lin proposed heuristics based method for computation of farthest pair of objects with $O(N)$ distance computation [89]. For an object x_1, the farthest object x_2 is computed and again x_3 is searched which is farthest from x_2. The steps are repeated a constant number of times to obtain the final pair, maintaining the linearity of heuristics. Vassilis *et al.* have proposed a random selection of N objects from the complete set and the generation of hash functions based on these objects [311]. In

the above heuristic methods, the certainty of dissimilarity of objects in a object pair is not guaranteed.

In the proposed approach for pivot object selection, we perform clustering of the training objects. Ideally, each cluster should have all the occurrences of an object in a single cluster. The clustering extracts the multi modal distribution information of the objects. We select these modes as the set of pivot objects where a cluster center is considered as the representation of mode. The clustering based selection identifies distinct points as pivot objects where each point will represent a group of similar objects. Additionally, the selection of cluster centres as pivot objects will ensure maximum spread of the distance between pivot objects. We compute the cluster center as the mean of data points belonging to a cluster, which minimizes the effect of noisy objects grouped wrongly in the clusters.

4.4.3 Locality Sensitivity Analysis of Distance based Hashing Functions

In the following discussion, we investigate the locality sensitivity of DBH with object representation defined in Euclidean space with Euclidean distance as the similarity measure. The discussion shows the applicability of a clustering based approach for pivot selection for locality sensitive hashing based on the Distance based hashing functions. This is not available in the literature to the best of our knowledge. The definition of LSH function requires knowledge about the underlying embedding of data points. Some of the references defining LSH functions are discussed in [141, 64, 44]. Aristides *et al.* have defined the

hashing function that embeds the data points in a Hamming cube [141]. In [64], p-stable distribution based hashing functions are defined as $f_{a,b}(q) = \lfloor \frac{a.q+b}{r} \rfloor$; p is d-dimensional input vector. The parameter a is a p-stable distributed d-dimensional random vector, and b is a uniformly distributed real random number between [0, r]. Charikar [44] has defined a random hyperplane based hash function which measures the probability of collision in terms of a defined similarity measure. The hash value represents the signum of projection of data points on the hyperplane. In contrast to LSH functions, the DBH functions perform object mapping without requiring knowledge of the object space geometry. The only requirement for hashing function definition is the existence of the distance measure $\mathcal{D}$ for the object space. However, the study of locality sensitivity of the $\mathbb{H}_{\mathbb{DBH}}$ requires complete geometric information of the object space.

The analysis of LSH directly is not applicable for evaluating $\mathbb{H}_{\mathbb{DBH}}$ because of the characteristics of equation (4.4.4). Considering two similar objects, represented by data points x_a and x_b. The similarity of these objects is defined by the close positioning of x_a and x_b. The projection of these points on the line joined by (x_1, x_2) is computed by equation (4.4.1),

$$\begin{aligned} & \{F(x_a) - F(x_b)\}^{x_1,x_2} \\ = & \frac{\mathcal{D}(x_1,x_a)^2 - \mathcal{D}(x_2,x_a)^2 - \mathcal{D}(x_1,x_b)^2 + \mathcal{D}(x_2,x_b)^2}{2 \times \mathcal{D}(x_1,x_2)} \\ = & \frac{(x_b - x_a)^T (x_1 - x_2)}{\mathcal{D}(x_1,x_2)} \end{aligned} \tag{4.4.5}$$

For the similar objects, i.e., $x_a \cong x_b$, the expression shows with good probability the projection by equation (4.4.1) will be *close*. For an uniformly distributed object data

space, the parameters $[t_1, t_2]$ estimated by (4.4.3) divide the line connected by (x_1, x_2) equally. In that case, for objects (x_a, x_b) having $close$ projection on the line connected by (x_1, x_2), the $\Pr(F(x_a)_{t_1,t_2}^{x_1,x_2} = F(x_b)_{t_1,t2}^{x_1,x_2})$ can be increased by improving the estimation of $[t_1, t_2]$. The estimation of parameters $[t_1, t_2]$ is done for the training set $\mathbb{X}$. It is evident that for given F, accuracy of $[t_1, t_2]$ depends on the size of $\mathbb{X}$, and on the spread of (x_1, x_2). Increasing the size of $\mathbb{X}$ and the selection of objects (x_1, x_2) such that maximum distance information is extracted, increases the probability of $F(x_a)_{t_1,t_2}^{x_1,x_2} = F(x_b)_{t_1,t2}^{x_1,x_2}$. Equation (4.4.1) is rewritten as

$$F^{x_1,x_2}(x) = \frac{(x_2 - x_1)^T(x - x_1)}{\mathcal{D}(x_1, x_2)} \tag{4.4.6}$$

The form of above equation is similar to random hyperplane based hashing function in [44]. The hashing function in [44] partitions the object space based on the signum of projection value on the randomly selected hyperplane from multivariate Gaussian $N(0, 1)$. In the present case equation (4.4.6) computes the inner product of random hyperplane $(x_2 - x_1)$ with input x, and partitions the object space based on (t_1, t_2). Rewriting the expression for collision probability as

$$\begin{aligned} &\Pr[F(x_a)_{t_1,t_2}^{x_1,x_2} = F(x_b)_{t_1,t_2}^{x_1,x_2}] \\ &= \Pr[\{(x_2 - x_1)^T x_a\}_{t'_1,t'_2} = \{(x_2 - x_1)^T x_b\}_{t'_1,t'_2}] \end{aligned} \tag{4.4.7}$$

In the simplification process, parameters t_1, t_2 are modified to t'_1, t'_2. Let us consider the special case of upper threshold $t'_2 \to \infty$ and lower threshold $t'_1 = 0$ in equation (4.4.7). In this case, we follow the result given Goemans and Williamson in [112] to compute the

probability in the (4.4.7).

$$\Pr[\text{sign}(x_2 - x_1)^T x_a \\ = \text{sign}(x_2 - x_1)^T x_b] = 1 - \frac{1}{\pi} \arccos \left\{ \frac{x_a x_b}{\|x_a\| \|x_b\|} \right\}$$

The special case discussed above is equivalent to signum of projection function in equation (4.4.6), which is locality sensitive. The clustering based approach for pivot object selection (x_1, x_2) helps in improving the accuracy of threshold parameters (t_1, t_2). In such case for the hashing function F defined over a large, uniformly distributed dataset $\mathbb{X}$ represented by data-points in Euclidean space, the estimate of t_1 is close to 0 and t_2 is a large number. The hashing performed by distance based hash function F therefore, will follow the locality preserving property.

4.4.4 Hierarchical DBH

In practice, most of the real datasets are non uniformly distributed. Though the DBH partitions the hash space considering the inter object distance distribution, it may lead to few densely populated and remaining sparsely populated buckets. This considerably reduces the efficiency of DBH in terms of accuracy and reduction gain in average comparisons for nearest neighbour search. The performance of several learning algorithms have been improved by maintaining hierarchy. We take this idea and build hierarchy of hash tables using data points in different buckets for improving the hashing performance (Figure 4.7). We can have maximum 2^k buckets in a hash table. In practice, the bucket count is much less in number, however; implementing hierarchy for all the buckets is not

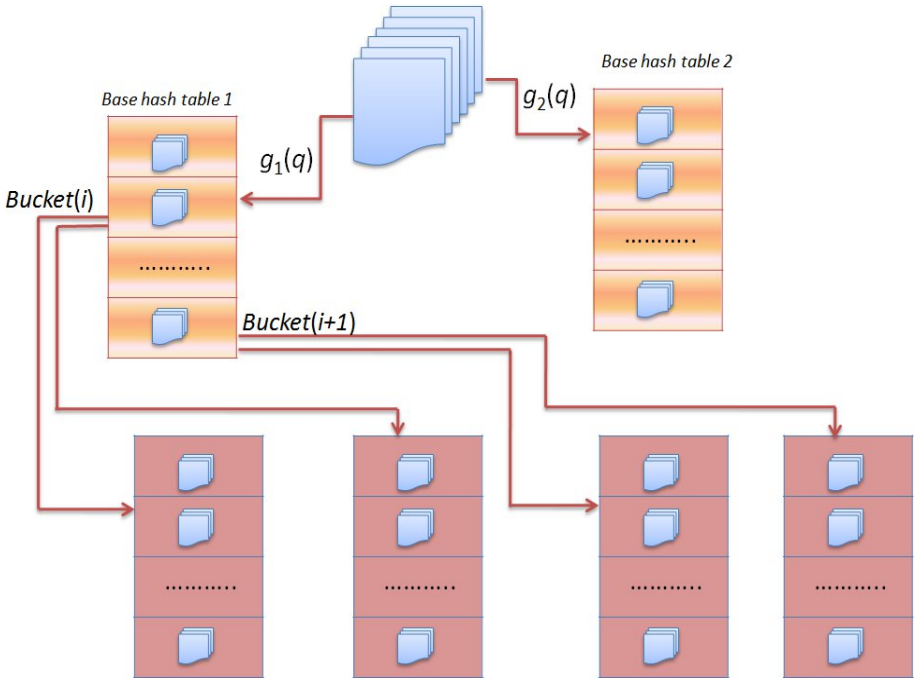


Figure 4.7: Hierarchical hash table generation

justifiable. Therefore, for hierarchical hash table generation, selection of buckets should follow certain criterion. The bucket selection for successive hashing can be based on either the population criterion or distribution information of objects in various buckets. We can consider constant or variable number of buckets for hierarchical hash table generation. The hierarchy generation process will be terminated after processing all selected buckets. In this case, hashing functions for hash table belonging to each bucket needs to be regenerated with objects hashed in the same bucket. In the proposed document indexing framework in section 4.5, buckets for rehashing have been selected based on population criterion. For DBH data structure with single hash table generated for k-bit hash functions, upper bound of retrieval time complexity will be $O(N + 1 - 2^k)$. In this case, the upper bound of retrieval time complexity for hierarchical DBH generated for most populated bucket, will

be $O(N + 2(1 - 2^k))$.

4.5 Experimental Results and Discussion

The experimental evaluation of proposed document indexing and retrieval framework is presented in this section. We have performed both experiments on three document collection of Devanagari, Bengali and English scripts. Devanagari and Bengali belong to Alpha-syllabic writing system and English belong to Alphabetic writing system. These scripts display great structural variation and varying combination of different constituents (vowels and consonants) in word formation. Additionally, Devanagari and Bengali scripts include long list of modifiers. In such scenario, word shape representation becomes difficult because of the complex composition of curved and straight character segments. The challenge is further compounded by varying typing styles. Therefore, it requires sufficient descriptor points to capture the shape characteristic and large set of angular and distance bins for accurate estimation of point-pair distribution. This increases the descriptor computation time and gives rise to high dimensional feature space. The high dimensional features incur high matching cost and increases memory storage requirement which is proportional to $O(mn)$. The concepts presented here are implemented in Matlab 7.6 environment. The simulations are performed on a 2GHz desktop computer with 1GB RAM.

Devanagari and Bengali document collection contains 503 pages scanned from 6 books and 226 pages scanned from 4 books respectively[1]. English document collection

[1]Document collections used for the experiments are sampled images from the dataset available at `http://ocr.cdacnoida.in`. The dataset is prepared as part of consortium research project

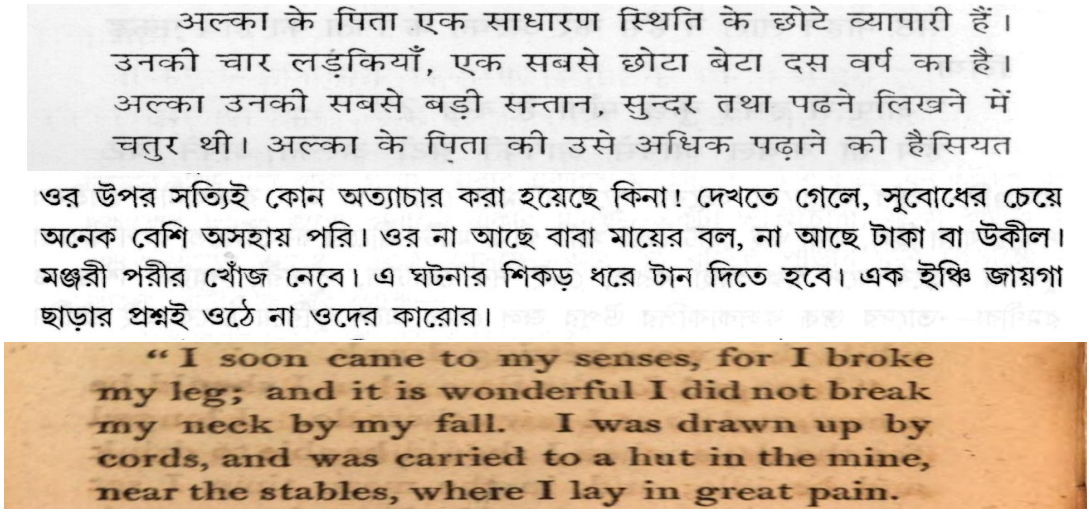


Figure 4.8: Sample document images

contains 212 pages from 6 books. The collection is compiled by sampling document images from the Google book dataset [116]. The dataset contains scanned images of old Latin script books. Sample of document images are shown in figure 4.8. The images from the collections are of low quality, primarily because of degradation in original document pages. The preprocessing steps for Devanagari and Bengali documents included smoothing and deskewing. English document images have been used in original form without any preprocessing. The conversion of original gray scale images to binary images is performed by Otsu's method. The word segmentation from the document images is done by horizontal and vertical profile based technique. After initial filtering, Devanagari word dataset contains 23145 words, Bengali word dataset consists 18632 words and the English word dataset consists of 19721 words. The filtering process removes stop words, punctuation marks, and words having length less than the defined threshold. The feature

funded by the Government of India.

representation discussed in section 4.3 is used for word image representation.

The document indexing framework for Devanagari, Bengali and English document collection is tested for 481, 278 and 301 queries respectively. The query words for Devanagari and Bengali document retrieval have 3 to 8 characters. In case of English documents, query word length varied from 4 to 11 characters. The retrieval experiment is performed for two categories of shape descriptors. In the first category, word image shape descriptors for different parameters (m, n) are computed without partitioning the image, i.e., the shape descriptor points are assumed belonging to single partition. In the second category, we split the word image in fixed number of partitions of uniform width, and compute the shape descriptor as discussed in section 4.3. The selection of number of partitions is based on heuristics that in Devanagari and Bengali scripts the maximum number words are formed with combinations of 3 to 5 characters. Therefore, for Devanagari and Bengali scripts word images we selected $1\times$ 4 partition for splitting the word image. Similarly most of the English words are formed combining 4 to 7 characters; therefore, two set of shape descriptors using 1×4 and 1×6 partitions are selected for splitting the word images. The high complexity in word shape requires significant number of descriptor points to capture the shape information. The initial evaluation showed that very high number of descriptor points incorporated noise in the set P (section 4.3). To avoid that, for all the experiments, the logical grid for point extraction is placed at interval of 4 pixels in both horizontal and vertical direction. The selection of descriptor parameters (m, n) is based on some preliminary observations. For very less number of bins, the uniqueness of point distribution histograms is lost. Therefore, we select sufficiently large number of angular and distance

bins such that the descriptor accurately captures the discriminative pattern of the point distribution histograms. The experimental evaluation showed that with increase in shape descriptor parameters, the discriminative ability of feature increases with increased word matching cost. However for very large number of bins (m, n), sparsely distributed point distribution histograms are highly noisy and very sensitive to the presence of different types of degradation and varying typing conditions in the document. The preliminary experiments are performed for precision oriented retrieval using nearest neighbour search. Precision oriented retrieval evaluates the result quality based on the precision of retrieved results with respect to the relevance to query. We selected five nearest neighbours for precision computation. Therefore, performance is measured as average precision of retrieved results in 5 nearest neighbours computed using Euclidean distance based similarity. For all the word datasets, the subset of query set having more than five similar examples are considered for evaluation. The precision with respect to descriptor parameter is presented in table 4.2.

Table 4.2: Precision oriented retrieval considering five nearest neighbours

	Devanagari	Bengali	English
Desc. paras. (m, n)	(45, 30)	(45, 30)	(50, 40)
Precision	90.76	91.54	89.86

In the experimental framework, the hierarchical hash tables are generated for two most populated buckets in base hash tables. The generation of hashing function family $\mathbb{H}$ is done using the cluster centres as pivot objects. The clustering over training set is performed by DBSCAN algorithm [86]. The search radius selection for computing the

precision and recall scores is done using the estimate of within cluster distances. The descriptor parameters (m, n) and hashing parameters (L, k) are the adjustable parameters. The implication of these parameters have been discussed earlier. A set of descriptor parameter values have been selected and the best results are presented. The best results therefore correspond to the optimal descriptor parameters for our document image collection. In retrieval experiments precision and recall have inverse relationship; therefore, F-score based single-point measure is considered for the selection of best. For Devanagari and Bengali document collection, shape descriptors are calculated with parameters (m, n) as $\{(45, 30), (45, 40), (50, 45)\}$ with 1×4 partition. Additionally, shape descriptor is also computed without partition i.e. global shape descriptor discussed in section 3.3.3. The descriptor computation steps include point extraction, *pdh* and Fourier transform computation. The time consumption for Fourier transform depends on $\{m, n, num\_parts\}$. Total computational time depends on the complexity of word shape with major part is spent on descriptor point extraction and *pdh* computation. Figure 4.9 shows the average descriptor computation time for Devanagari word collection with respect to different descriptor parameters.

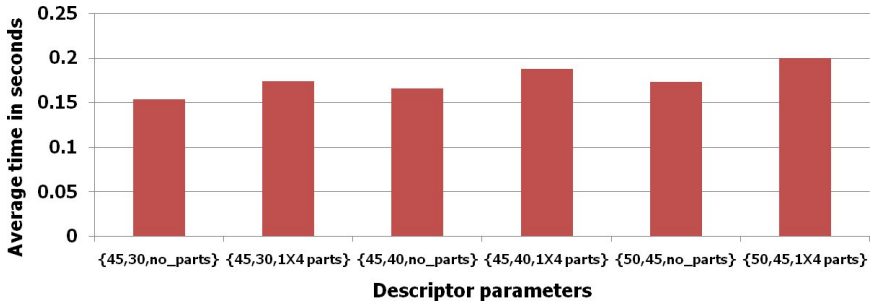


Figure 4.9: Computation time for descriptor computation for Devanagari words

Table 4.3: Devanagari retrieval results for descriptor parameters $\{m = 50, n = 45\}$: without partition and with 1×4 partition

Hashing	$L = 15, k = 15$		$L = 15, k = 18$		$L = 18, k = 20$		OCR based
Paras.	No Part.	1× 4 Part.	No Part.	1× 4 Part.	No Part.	1× 4 Part.	results
Precision	86.45	88.18	86.90	88.42	87.07	88.64	82.84
Recall	87.26	88.79	84.25	85.93	83.44	85.18	81.55
F Score	86.85	88.44	85.55	87.15	85.22	86.88	82.19
Avg. Comp.	2993	2775	2295	2088	2163	1977	
Avg. time	0.67	0.66	0.58	0.55	0.55	0.49	

Table 4.4: Bengali retrieval results for descriptor parameter $\{m = 50, n = 45\}$: without partition and with 1×4 partition

Hashing Paras.	$L = 15, k = 15$		$L = 15, k = 18$		$L = 18, k = 20$	
	No Part.	1× 4 Part.	No Part.	1× 4 Part.	No Part.	1× 4 Part.
Precision	86.85	88.52	86.99	88.67	87.16	88.95
Recall	87.81	89.45	85.04	86.60	84.27	85.77
F Score	87.33	88.98	86.00	87.62	85.69	87.33
Avg. Comp.	2374	2139	1862	1695	1723	1579
Avg. time	0.54	0.52	0.46	0.43	0.44	0.42

The empirical observation showed that for the same hashing parameter, descriptor parameters ($m = 50$,$n = 45$) are optimal for Devanagari and Bengali word images. The retrieval results corresponding to these parameters are presented in table 4.3 and 4.4. For the same set of descriptor parameter, LSH based retrieval results on Devanagari word collection is also presented in table 4.5. To define LSH based indexing, random hyperplane based hashing functions have been applied [44]. The comparison between table 4.3 and 4.5 showed that DBH based indexing achieved 3.36∼4.65% improvement in F-score for

different hashing parameters. The DBH achieved better precision and recall compared to LSH requiring less number of average computations for larger k. It is justified as random hyperplane based projections are independent of data distribution; therefore, for smaller k, the hashing function g shows poor discriminative power. However, in practice, short hash functions (small value of parameter k) are preferred at acceptable retrieval performance to control the size of indexing data structure. In this context our results are in accordance with recent results presented by Muja and Lowe [221]. The authors have empirically shown that hierarchical clustering trees based on randomized KD tree or hierarchical k-means achieved better or comparable performance than LSH in terms of precision, speed and storage requirement for large scale search.

Table 4.5: LSH based retrieval for Devanagari collection with descriptor parameters $\{m = 50, n = 45\}$: without partition and with 1×4 partition

Hashing Paras.	$L = 15, k = 15$		$L = 15, k = 18$		$L = 18, k = 20$	
	No Part.	$1\times$ 4 Part.	No Part.	$1\times$ 4 Part.	No Part.	$1\times$ 4 Part.
Precision	81.78	83.31	83.11	84.55	83.63	84.97
Recall	83.19	84.27	80.54	81.75	80.19	81.36
F Score	82.48	83.79	81.80	83.13	81.87	83.13
Avg. Comp.	3425	3289	1897	1724	1648	1489
Avg. time	0.75	0.73	0.47	0.43	0.42	0.40

The proposed indexing and retrieval framework is also validated with respect to retrieval performance of a search engine which use recognized characters OCR'ed in the document image. The experiment is performed on Devanagari document images. The recognition is performed by Devanagari OCR discussed in [12]. The character level

recognition accuracy of 86.56% is achieved. Figure 4.10 shows sample document image from the collection and the recognized text. For the same query set, precision and recall rate of 82.84% and 81.55% is achieved. Table 4.3 shows that proposed framework achieves best F-score of 88.44% with precision and recall as 88.18% and 88.79% respectively. The blue boxes in the left-side images show the retrieved word image for query word साक्षात्कार, and red characters/symbols in right-side images are the wrong recognitions. The degradation in document images causes poor recognition. Evidently, the unicode based indexing of OCR'ed image skips the words even with single recognition error which reduces the retrieval performance significantly. In this case, the query retrieval time varied

Figure 4.10: Sample document images and corresponding OCR'ed output

from 0.49∼0.67 seconds. The Fourier transform computation is time consuming; however, the computation is performed once during on-line querying. Additionally, the DBH based approximate nearest neighbour search significantly reduces the search complexity as shown

by approximation ratio of 8.5$\sim$12.9% in average computations.

Following the similar evaluation methodology, the indexing framework is applied for English script documents. A set of descriptor parameters (m, n) are selected as $\{(50, 45)$, $(38, 36)\}$ without and 1×4, 1×6 partition respectively. Based on the F-score, $(m = 38, n = 36)$ with 1×6 partition achieved best result for the selected hashing parameters. The corresponding retrieval results are presented in table 4.6. The retrieval performance

Table 4.6: English retrieval results for $\{m = 38, n = 36\}$: without partition and with 1×6 partition

Hashing Paras.	$L = 15, k = 15$		$L = 15, k = 18$		$L = 18, k = 20$	
	No Part.	1× 6 Part.	No Part.	1× 6 Part.	No Part.	1× 6 Part.
Precision	83.78	86.24	83.96	86.39	84.14	86.64
Recall	84.91	87.95	81.02	85.98	80.12	84.86
F Score	84.34	87.08	82.46	86.18	82.08	85.74
Avg. Comp.	3397	2716	2487	2041	2098	1903
Avg. time	0.79	0.66	0.59	0.54	0.54	0.53

over the English script documents is compared with the word shape code based retrieval presented by Lu *et al.* in [193]. On our experimental dataset with similar set of queries, word shape coding based approach achieved precision and recall rate of 82.47% and 78.54%. In this case, the shape descriptor representation achieves best F-score of 87.08% with precision and recall as 86.24% and 87.95%. The morphological operations applied for word code generation in [193] are sensitive to noise and document degradations. Therefore, the method requires application of strong document enhancement technique in case of old and degraded documents. Additionally, shape code based approach is not extendible to

other scripts.

Next, the proposed framework is evaluated for indexing a synthetic document image dataset prepared by Reuter-21578 text collection. The objective is to evaluate our framework with the method presented in [193] by performing retrieval on synthetically created document images. The dataset generation and evaluation strategy is followed by [193]. The results in table 4.6 showed the descriptor parameters {$m = 38$, $n = 36$, 1×6 Partition} achieved best results. Therefore, these parameters are selected for word image representation. The word image segmentation is done following the projection profile based strategy. The words having less than 3 characters are filtered out in the preprocessing stage. The filtered word image collection consists 26700 samples. The query set consists of 125 frequently used word. The retrieval performance with the proposed indexing framework for different hashing parameters is presented in the table 4.7. In this case, the method presented in [193] achieved precision, recall and F-score of 94.12%, 91.86% and 92.98% respectively. Table 4.7 shows that the proposed indexing framework achieved comparable performance. The shape descriptor based representation is robust in case word shape deformations. The results establish that proposed framework provides an efficient solution for indexing the old documents where the degradations are stochastic in nature. Additionally, the framework provides the option of adjustable parameters which can be tuned according to the performance requirement.

The experimental observation showed improvement in precision with increase in k, though sharp decrease in the collision probability decreases recall rate significantly. Therefore, multiple hash tables are required to improve recall. Using word shape coding

Table 4.7: English retrieval results: Synthetic dataset prepared by Reuter-21578 text collection, descriptor computation with $\{m = 38, n = 36, 1\times 6$ Partition$\}$

Hashing Paras. (L, k)	(18, 20)	(18, 24)	(24, 24)
Precision	87.75	95.74	93.26
Recall	93.42	88.16	92.20
F Score	90.50	91.79	92.72
Avg. Comp.	4363	2686	3237
Avg. time	0.92	0.61	0.74

scheme for English words, Bai *et al.* [20] have reported F-score of 0.926 for a decent quality document collection. The document images used for evaluation of the proposed framework are low quality and contain significant amount of noise (Figure 4.8). However, with the original images for English script, best F-score of 0.87 is achieved. The variations in parameters L and k changes precision and recall. Nevertheless current hashing parameters are selected to achieve satisfactory F-score at reasonable search time complexity. The experimental results show that hierarchical DBH achieves significant improvement in search time complexity with approximation ratios of 9.6$\sim$17.2%. It is important to point that the presented results are based on single query word which retrieve the documents based feature based similarity of query to textual content of the document image. In case of query having multiple words, advance techniques needs to be explored in the direction of presented work. In this sense, the presented retrieval model does not completely follow an information theoretic approach. The direction of possible extension could be ranking based fusion for combining documents retrieved for each query.

4.6 Multi Probe Hashing in DBH Framework

The following discussion presents a new approach for Multi probe hashing in DBH framework. In [194], Qin Lv *et al.* presented the concept of Multi probe hashing which aims at reducing the size of LSH data-structure. In the conventional LSH, single bucket from each hash table is probed during retrieval. Instead of probing single bucket from each hash table for retrieval, the method intelligently selects more than one bucket which are likely to have similar objects to query and uses them for probing. The process subsequently reduces the requirement of large number of hash tables for achieving desired recall. Applying the LSH preamble, it is highly possible that similar objects are hashed in nearby, i.e., *adjacent* buckets. Therefore, searching the buckets *adjacent* to query bucket in a hash table can retrieve similar objects. The *adjacent* buckets are identified by applying a perturbation to the query bucket $g(q)$. The perturbation is defined as vector $\Delta = \{\delta_1, \delta_2, ..., \delta_k\}$, k is the length of function g. The δ_i define the perturbation corresponding to function f_i. Qin Lv *et al.* proposed two different methods for Multi probe hashing [194]. The first method defines an step wise probing; i.e., the address for new buckets to probe are generated by applying perturbation over s hash values in $g(q)$ vector. Therefore, for k length hashing function using s-step probing, we can generate $2^s \times^k C_s$ perturbation vectors and therefore, $2^s \times^k C_s$ buckets. The perturbation values δ come from $\{+1, -1\}$. The second method defines query directed multi-probing where the perturbation vectors to $g(q)$ are generated based on their estimated query dependent scores. In practice, it is always desirable to have binary hashing functions. However, the existing methods for multi-probe hashing

are for real hashing functions. The Distance based hashing functions are binary in nature; therefore, the direct extension of existing methods for DBH is not possible.

In the following discussion, we present a novel approach for Multi probe hashing using Distance based hashing functions. The step-wise perturbation has the advantage of giving equal importance to all the hashing functions $\{f_1, f_2, ..., f_k\}$ in g. Therefore, we follow step wise perturbation for identifying the buckets for multi-probing.

4.6.1 Step-wise Multi-probing in Distance Based Hashing

The application of step wise perturbation at s places will invert the hash value at s coordinates of $g(q)$. Therefore, we can identify *adjacent* buckets to the query bucket $g(q)$ whose indices differ at s coordinates. The set of bucket addresses obtained after perturbation are neighbourhood buckets to query bucket. We can select buckets for successive probing from this set. To apply s-step perturbation to $g(q)$, we generate ${}^kC_s$ k-bit vectors such that each vector has s 1's and remaining 0's. The basic idea is to apply perturbations to s sides of the boundary of query bucket. This can be achieved by XORing the query bucket address with perturbation vectors. Therefore, we will get maximum ${}^kC_s$ valid bucket addresses. Considering 4-bit query bucket address 1101, the 1-step perturbation vectors will be 1000, 0100, 0010 and 0001 and the probable buckets in which the objects similar to query may be hashed are 0101, 1001, 1111 and 1100. Similarly for 2-step perturbation, the set of perturbation vectors will be 1100, 1010, 1001, 0110, 0101 and 0011. Therefore, the probable bucket will be 0001, 0111, 0100, 1011, 1000 and 1110. The steps to generate step-wise perturbation vector set is defined below

(i) Enter k - length of query bucket address and perturbation step s.

(ii) Generate ${}^kC_s$ k-bit perturbation vectors with each vector having s 1's and remaining 0's.

The set of bucket addresses are obtained by following steps

(i) Enter k-bit query bucket address.

(ii) Generate ${}^kC_s$ bucket addresses. The i'th bucket address is obtained by XOR operation between query address $g(q)$ with i'th perturbation vector.

We select only valid bucket addresses from the generated set. These buckets represent the *adjacent* buckets to query bucket. The probability of hashing of similar objects to query q, in any of these buckets will be equal. We can utilize the population density information of each valid bucket to finalize the buckets for multi probing. Ideally, each bucket in a hash table is dominated by group of similar objects; therefore, alternately we can finalize the buckets for multi probing by ranking them based on the distance between query q from the center of valid buckets. In this case, the mean of data points representing objects in a bucket can be considered as bucket center.

4.6.2 Success Probability Estimation

The DBH function (4.4.2) does the object projection onto a line which is uniformly bi-partitioned by parameters (t_1, t_2). Let q be the query object and p is one of the nearest neighbour. Let μ be the probability of $f_i(p) = f_i(q)$, for i = 1, ..., k. The probability of q

to be hashed in adjacent bucket to $g(p)$ is

$$\Pr[g(p) = g(q) \oplus \Delta] = \prod_{i=1}^{k} \Pr[f_i(p) = f_i(q) \oplus \delta_i] \tag{4.6.1}$$

Here Δ = {δ_1, δ_2, ..., δ_k} is the perturbation vector with $\delta_i \in \{0, 1\}$. Since for s-step perturbation s δ_i's are 1's and remaining are 0's, the likelihood equation presented above can be written as

$$\begin{aligned}
&\prod_{i=1}^{k} \Pr[f_i(p) = f_i(q) \oplus \delta_i] \\
&= \prod_{i=1,\delta_i=0}^{k} \Pr[f_i(p) = f_i(q)] \prod_{i=1,\delta_i=1}^{k} \Pr[f_i(p) \neq f_i(q)]
\end{aligned}$$

For s-step perturbation, we can simplify the above equation,

$$\prod_{i=1}^{k} \Pr[f_i(p) = f_i(q) \oplus \delta_i] = \mu^{k-s}(1-\mu)^s \tag{4.6.2}$$

Equation (4.6.2) represents the probability of p hashed in *adjacent* bucket to $g(q)$, where the indices of *adjacent* buckets differ at s coordinates from $g(q)$. The probability is higher for small values of s which is acceptable since the bucket indices which differ from $g(q)$ at less coordinates values are more natural candidates for *adjacent* buckets. Even with high probability of collision μ, the success probability mentioned in equation (4.6.2) will be low in measure. The proposed approach can also be used for defining Multi probe hashing using other binary mapping functions. In the conventional hashing, the retrieval success rate is increased by pooling neighbours from multiple tables for performing similarity search. We can apply the same idea for increasing the $\Pr[g(p) = g(q) \oplus \Delta]$ by selecting more buckets instead of single bucket (final bucket to be probed)

from the addresses generated by step discussed in section 4.6.1, e.g. for 4-bit address 1101, we can select more than one buckets from 0101, 1001, 1111 and 1100 for multi-probing. Here, we observe that the objective of random perturbation based multi probing is to reduce the number of hash tables, whereas hierarchical hashing aims to reduce the average retrieval time. In both the cases, trade off has to be accepted in terms of accuracy and retrieval processing time.

4.6.3 Performance Evaluation

The performance evaluation of the multi probe hashing is done following the approach proposed in [194]. The experiments were performed on the document collection described in the section 4.5. The true percentage of K-NN's retrieved results, i.e., recall is considered as performance measure. In addition, average number of comparisons also present an estimate of performance improvement in terms of search complexity. Therefore, recall represents the quality factor and number of comparisons represents the cost factor. Considering q as query object and $I(q)$ are the K nearest neighbours representing ideal result. Suppose $A(q)$) are the K nearest neighbours obtained from the multi-probe hashing scheme. The recall is defined as

$$Recall = \frac{A(q) \cap I(q)}{I(q)} \tag{4.6.3}$$

Here, precision and recall are same because the retrieved objects (collection of objects obtained from the queried buckets) are ranked based on their similarity with the query and. We consider 10 nearest objects, i.e., K = 10 for measurement. The evaluation of

multi-probing in DBH is done on Devanagari, Bengali and English word image collection (details of the datasets are presented in the section 4.5). The recall for Devanagari dataset is computed for 481 query words and for Bengali with 278 words. The word length for the query words varied from 3 to 8 characters. The recall for English dataset is computed for 301 query words with word length varying from 4 to 11 characters. The experiment is performed for different values of L with only base hash table without considering hierarchical hash tables. The bucket selection for the multi-probing is done based on the population criterion from the addresses generated from steps discussed in section 4.6.1. We have performed Multi probe hashing using 1-step and 1,2-step perturbation. In 1,2-step perturbation, we combine buckets obtained using 1 and 2-step probing for similarity search.

The results in section 4.5 show that descriptor parameters $\{m$ = 50, n = 45, $1\times$ 4$\}$ are optimal for Devanagari and Bengali script words because of excellent discriminating characteristics across different words. Similarly for English words, descriptor parameter $\{m$ = 38, n = 36, $1\times$ 6$\}$ is the optimal. The multi-probe hashing results on the different script document collection is presented for these descriptor parameters. The results are presented in figures 4.11, 4.12 and 4.13. The recall score obtained by complete set of hash tables is compared with the recall score of Multi probe hashing considering $1/5^{th}$ of initial set of tables. The results show the recall score of 1,2-step multi-probing matches closely with recall obtained by complete set of tables. In most cases, the difference between recall scores varies from 0.30% to 0.90%. However, the experiments show minor increment in number of comparisons and average processing time which is primarily because of bucket

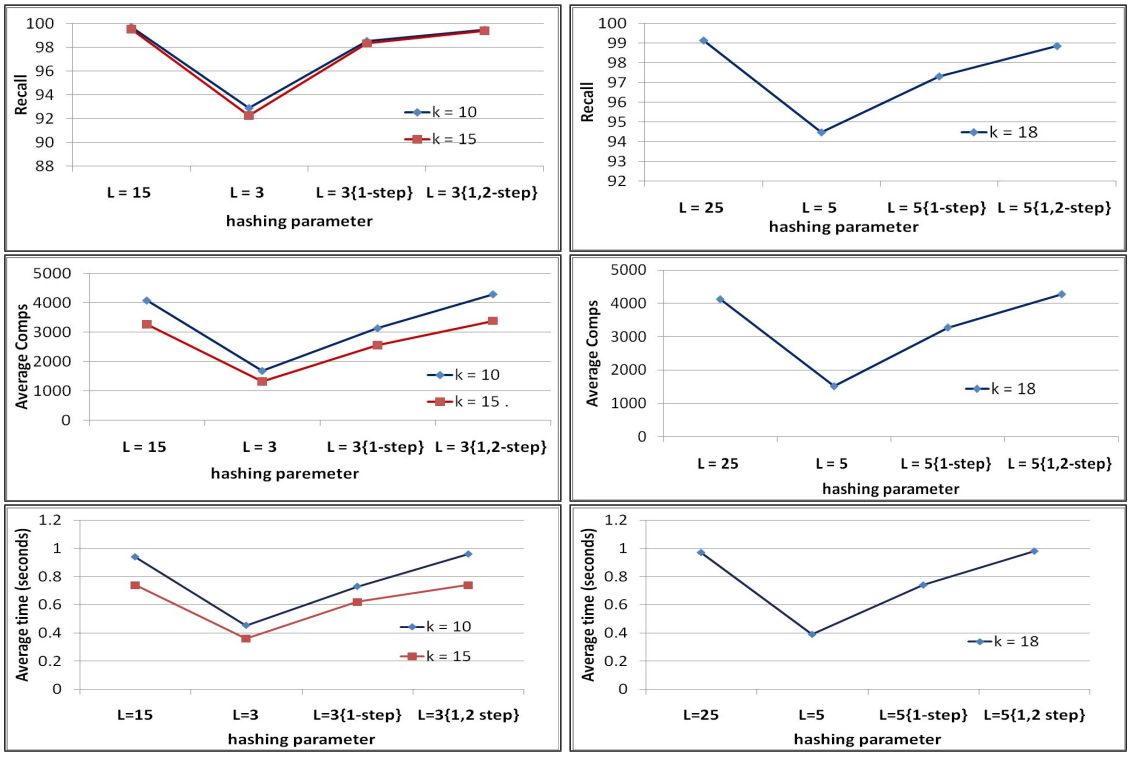


Figure 4.11: Multi-probe hashing results with Devanagari word dataset: $\{m = 50, n =$ 45,1×4 Partition$\}$

selection based on population criterion.

4.7 String like Word Representation for Document Image Indexing

section 4.3 presented shape based feature representation which encodes the global boundary information of word object in a real vector. The representation posses excellent discriminative ability across different class of word shapes. However, the feature representation

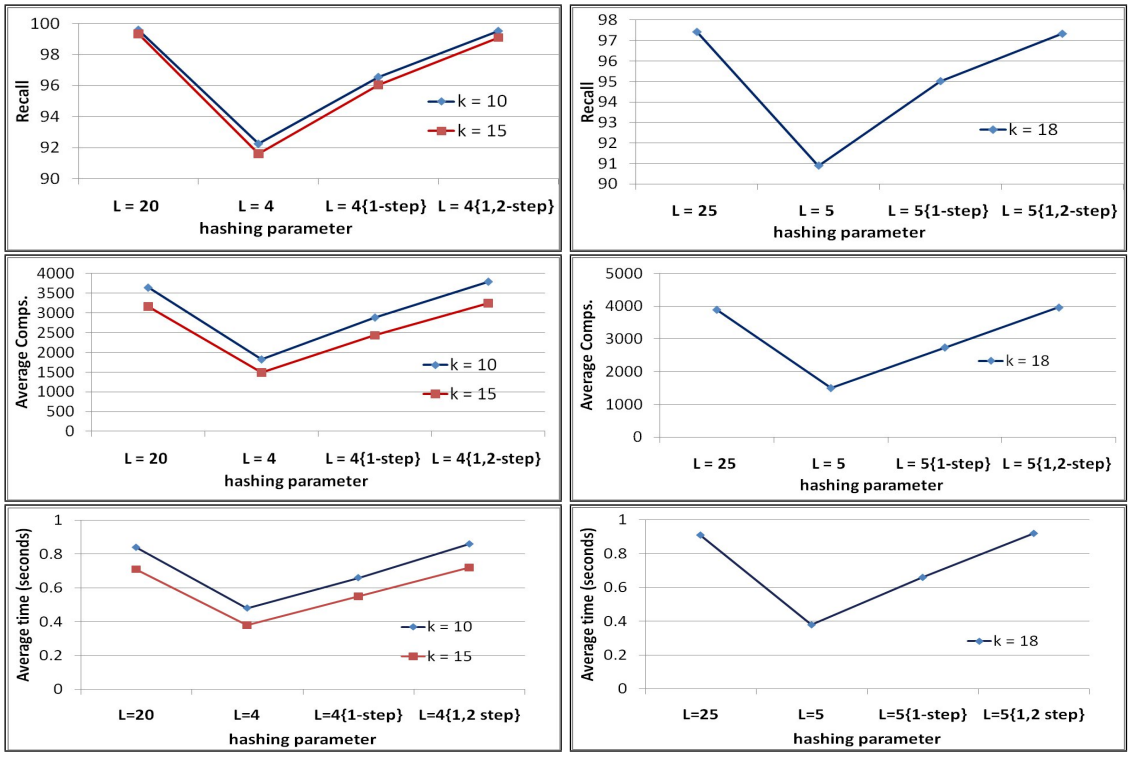


Figure 4.12: Multi-probe hashing results with Bengali word dataset : $\{m = 50, n = 45$,1×4 Partition$\}$

based retrieval results does not include sub-string matches to the query. For example the similarity matching using shape descriptor based representation would retrieve word भारतीय for query भारतीय without including अभारतीय. Similarly, for query इतिहास, the word इतिहास would match, leaving इतिहासकार. In this section, we present novel string based representation for word images. The representation is subsequently applied for developing document indexing scheme using edit distance based hashing. Novel clustering based method is developed for word image representation. We identify the graphical primitives used in word formation and apply adaptive clustering for grouping these primitives. A graphical primitive is the structural part of word image obtained after segmentation. The clustering

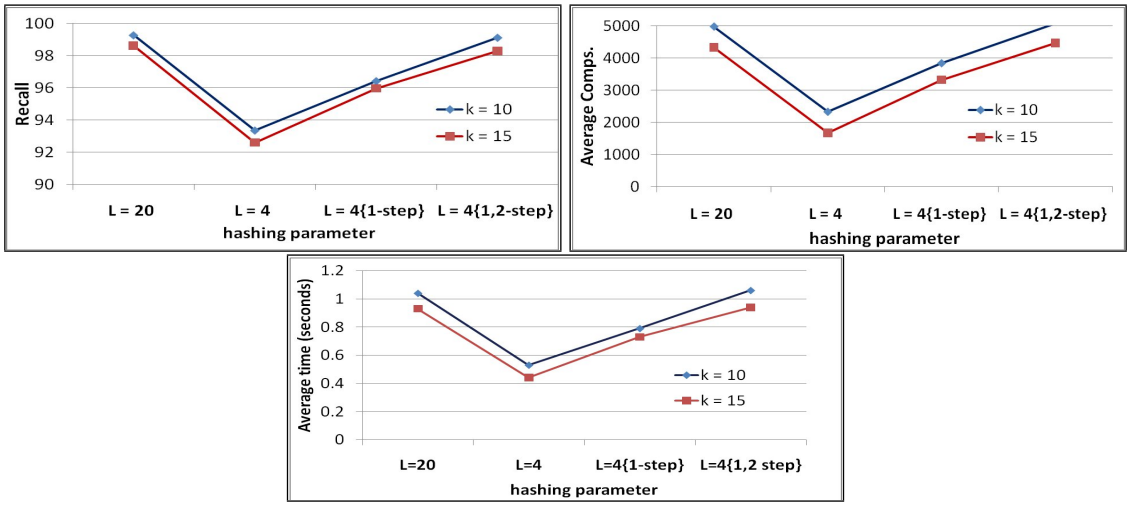


Figure 4.13: Multi-probe hashing results with English word dataset: $\{m = 38, n = 36$,1×6 Partition$\}$

identifies equivalence groups of graphical primitives existing in the collection of documents. Each group of graphical primitives is assigned an unique code. The word image representation is defined by identifying the graphical primitives, and assigning the code based on nearness to equivalence groups. In this context, the approach presents a script invariant methodology for word image representation.

4.7.1 Word Image Representation

We follow the string codes based representation for word images. Section 4.3.1 presented the issues related with the extraction of character primitives in Indian script words. Therefore, we define word representation based on graphical primitives instead of character primitives for string code generation which enhances the robustness of the representation. The segmentation of graphical primitives is relatively much simpler and it may have a

single letter, or combined letters defined by various grammatical rules. The segmentation of character primitives from word images are performed by identifying the local minima over the vertical profile of the word image. The cut-off points in figure 4.14 are the local

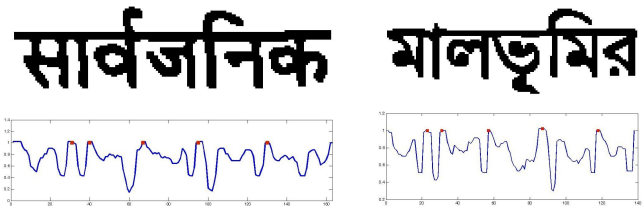


Figure 4.14: Vertical profile and cut-off points over for graphical primitive segmentation

minima points. Each graphical primitive is defined by the region between cut-off points including the end points. The process generates a large set of graphical primitives. The figure 4.15 shows set of the graphical primitives obtained after segmentation. The next step

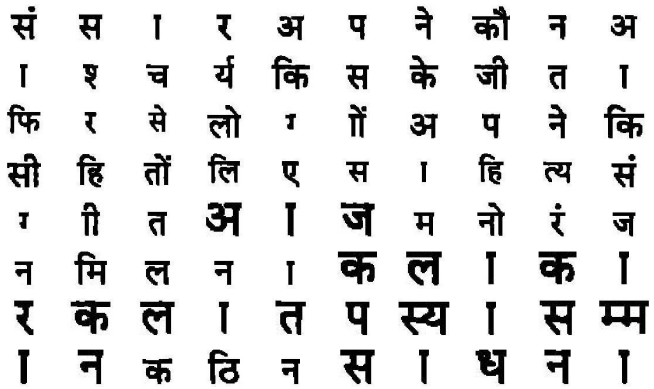


Figure 4.15: Sample graphical primitives from Devanagari document collection

for word representation is code assignment for each graphical primitive. In the context of Indian scripts, the exact estimation of unique graphical primitives is a difficult task. The problem is further compounded by noisy character primitives segmented wrongly because

of the degradation in document images. To increase the robustness of the character code, we follow clustering based approach for learning the codes. A sample set of words is chosen to generate the character primitive dictionary. The words are selected such that segmented primitives cover maximum range in terms of variety. Since the prior knowledge about the dictionary size is not available, we apply DBSCAN for clustering [86]. Each character primitive is assigned to a cluster with respect to its nearness to the cluster. The nearness is computed by Euclidean distance between cluster primitive and cluster centres.

$$\texttt{String}(W) = [\texttt{Gr}\{Pr_1\} : ... : \texttt{Gr}\{Pr_m\}]$$

Here $\texttt{Gr}(Pr_i) = \underset{j=1,...,\texttt{no\_Eq\_groups}}{\text{argmin}} \{\|Pr_i - \texttt{Eq\_group}_j\|\}$.

4.7.2 Document Indexing using Edit distance based hashing

The document indexing using the string code for word image representation follows the conventional hashing based indexing framework discusses in section 4.2. Equation (4.4.1) shows that availability of distance $\mathcal{D}$ is the only requirement for projection. Therefore, the mapping F is also applicable for arbitrary spaces. Here, we define edit distance based hashing functions for generating the word image indices. The edit distance computes the similarity between two strings as minimum number of edit operations required to transform one string to other. In this case, set of edit operations include insertion, deletion, or substitution of single character. Mathematically, the edit distance between string s_1, and s_2 defined as, $ED_{s_1,s_2}(|s_1|, |s_2|)$ is computed as:

$ED_{s_1,s_2}(i,j)$ =

$$\begin{cases} 0 & \text{If } i=j=0 \\ i & \text{If } j=0 \text{ and } i>0 \\ j & \text{If } i=0 \text{ and } j>0 \\ \min \begin{cases} ED_{s_1,s_2}(i,j-1)+1 & \texttt{//Insertion} \\ ED_{s_1,s_2}(i-1,j)+1 & \texttt{//Detection} \\ ED_{s_1,s_2}(i-1,j-1)+[s_1(i)\neq s_2(j)] & \texttt{//Substitution} \end{cases} & \text{Otherwise} \end{cases}$$

Therefore, the edit distance based word image indexing in hash space should group words having completely and partially similar string representations.

4.7.3 Experimental Evaluation

The proposed document indexing framework is evaluated on document image collection of Devanagari and Bengali script discussed in section 4.5. The preprocessing and word image segmentation procedure is followed as discussed earlier. Conventional performance measures i.e. Precision and Recall are computed over unordered result set. However, the ranking of retrieved results is an important retrieval measure. Therefore we computed mean average precision (MAP) for measuring the performance of proposed indexing scheme. The MAP for a query set X_v is computed as the mean of average precision values [200].

$$\text{MAP}(X_v) = \frac{1}{|X_v|}\sum_{j=1}^{|X_v|}\frac{1}{m_j}\sum_{k=1}^{m_j} Precision(R_{jk}) \tag{4.7.1}$$

Average precision for query $q \in X_v$ is defined as mean of precision at each relevant recall point in the retrieved results. In equation (4.7.1), m_j represents the number of relevant retrieved results for query q_j and R_{jk} represents the ranked retrieval results from the top to k^{th} relevant result. If a retrieved document in R_{jk} is non-relevant, the precision value

at this recall point is not considered. In the present indexing scheme, average precision for q is computed over the collection of neighbours obtained from L hash tables having indices $g_i(q)$ for $i = 1, ..., L$. The neighbours are ranked based on the Edit distance from the query string q. The Edit distance for word matching is computed as Levenshtein algorithm between two strings.

The shape descriptor for Devanagari graphical primitives are computed for (m = 40, n = 30). For similar hashing parameters, we have also computed the indexing performance with shape descriptor representation for word images as discussed in section 4.3. The word image similarity is established by Euclidean distance. The descriptor computation is performed for {m = 40, n = 30, 1×4 partition}. For Bengali graphical primitives, the shape descriptor is computed for m = 40, n = 36). For shape descriptor representation based word images, the following parameters are is selected {m = 40, n = 36, 1×4 partition}. The MAP and average comparisons for different hashing parameters (L, k) are presented in the table 4.8. The evaluation shows comparable performance of string

Table 4.8: Document retrieval results with edit distance based hashing

Results with Devanagari documents				
Descriptor type	String codes for word images		Shape descriptor for word images	
Hashing paras.	MAP	Comps.	MAP	Comps.
L = 50, k = 12	71.56	1445	76.12	1729
L = 50, k = 20	70.09	1113	72.08	1212
Results with Bengali documents				
Descriptor type	String codes for word images		Shape descriptor for word images	
Hashing paras.	MAP	Comps.	MAP	Comps.
L = 50, k = 12	74.23	1286	78.65	1517
L = 50, k = 15	68.70	712	71.26	879

based word representation with edit distance based indexing with the shape descriptor based word representation using Euclidean distance based indexing. The average number of comparisons of significantly less. However, the word object grouping in hash space using edit distance based similarity is not comparably accurate as shown by the MAP measures. Nevertheless, the scheme provides a satisfactory alternative to shape descriptor based distance based indexing as it requires less disk space for storage, and also considers sub-string matches in retrieved results.

4.8 Conclusions

A novel word image based document indexing scheme is proposed which applies the concept of distance based hashing for indexing. We presented a novel feature based indexing scheme for Indian script documents. The experimental evaluation of the indexing framework is performed on Devanagari, Bengali and English script document collections. Despite the poor image qualities of experimental documents, the framework achieves F-scores of 0.88, 0.89 and 0.87 for Devanagari, Bengali and English scripts respectively. We also presented new method for multi probing in the case binary mapping functions. In addition, novel word string based document indexing framework is presented using edit distance based hashing. We explore the applicability of distance based hashing for defining multiple feature based document indexing in the chapter 5.

Chapter 5

Learning for Document Image Indexing with Multiple Features

5.1 Introduction

Indexing is a key requirement for designing multimedia retrieval systems. Fundamentally, Indexing schemes target efficient retrieval of the set of objects similar to the query based on predefined ranking mechanism. The chapter concentrates on the development of feature based indexing mechanism for image collections. Currently, feature based indexing schemes have been used in different applications [78, 262, 356, 163, 189, 288]. Recent research in computer vision have shown improved results in many problems by applying kernel function based similarity metric. In this work, a novel image indexing scheme based on Kernel distance based hashing (KernelDBH) is presented. Kernel functions can directly model the similarities between samples in different feature spaces. We present the

novel extension of distance based hashing to kernel space which generates the indexing structure based on kernel distance based similarity.

A novel for integrating multiple features for defining the composite/different indexing space is proposed by applying the concept of Multiple kernel learning (MKL). The existing MKL algorithms developed for classification problems learn the combination of features while searching for maximum margin boundary planes by solving a joint optimization problem. However, Multiple kernel learning for Indexing requires optimization of retrieval performance as the objective. In this work, the distinct nature of indexing objective is addressed by defining a different MKL formulation. The retrieval performance based objective makes the conventional optimization techniques inapplicable. In this direction, we propose application of Genetic algorithm (GA) for performing the optimization task in MKL.

We have evaluated the proposed scheme for a bench mark dataset of handwritten digit images. Further, using the same scheme we have developed word image based document image retrieval application. Here, we propose set of features for word image by exploiting the distinct word shape properties. The indexing scheme applies the features individually, and by combination to create the indexing data structure. The experimental results are presented on Devanagari, Bengali and English document image collection. In summary, the major contributions are as follows.

- Image indexing scheme using KernelDBH by object index generation based on kernel distance similarity.

- MKL formulation in indexing framework to learn the optimal kernel for KernelDBH. The framework provides a principled and logical approach to apply multiple features in indexing applications.
- Application of GA for the optimization task in MKL.
- Development of word based document image indexing using the proposed indexing scheme for Devanagari, Bengali and English script documents.
- In addition, the generalization of proposed concept is shown by evaluation on handwritten digit and natural image collection.

5.2 Distance based Hashing in Kernel Space

In the following discussion, the extension of distance based hashing to kernel space is presented for developing the proposed image indexing scheme using multiple features. The details of distance based hashing have been discussed in section 4.4.1. Figure 5.1 shows overall framework for indexing the image collection using distance based hashing functions.

5.2.1 Proposed Kernel based DBH

In kernel based learning methods, the kernel matrix represents object similarities in high dimensional feature space. The transformation to higher dimensional space, i.e., kernel space, by *kernel trick* helps to extract the similarity information between high dimensional data points which otherwise is difficult to extract in input space. The application of

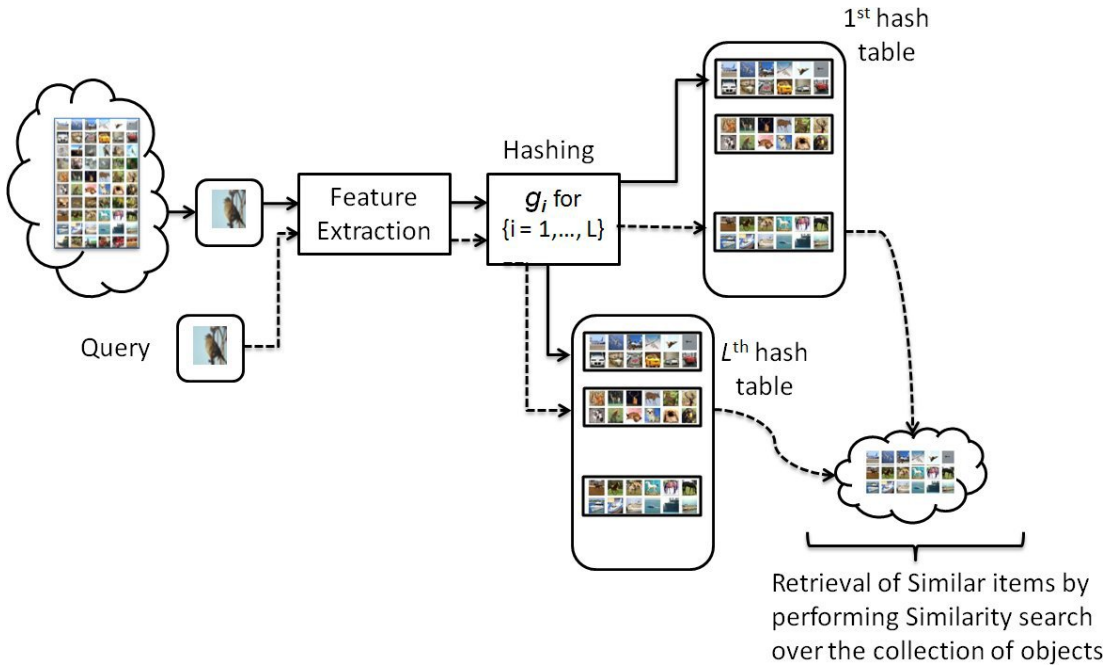


Figure 5.1: Indexing scheme using distance based hashing. The dotted lines show the query retrieval process.

kernel matrix has been the preferred data similarity measure for various computer vision problems. In this section, we propose extension of DBH to Kernel based DBH. The extension provides the platform to use kernel matrix as the distance measure for indexing. However, the fundamental property of DBH i.e. distance preservation in hash space is maintained. Considering $\mathcal{X}$ as Euclidean vector space and $\mathcal{D}$ Euclidean distance, the squared distance $\mathcal{D}^2(x_1, x_2)$ can be expanded as $x_1^T x_1 + x_2^T x_2 - 2x_1^T x_2$. Equation (4.4.1) is redefined as

$$F^{x_1,x_2}(x) = \frac{x_1^T x_1 - x_1^T x + x_2^T x - x_1^T x_2}{\sqrt{x_1^T x_1 - 2x_1^T x_2 + x_2^T x_2}} \tag{5.2.1}$$

The above expression represents the line projection computation using dot products. The kernel methods increase the computational power of linear learning algorithms by mapping the data to potentially higher dimensional feature space [271]. Let ϕ be the nonlinear mapping which does the transformation from input space $\mathcal{X}$ to high dimensional space (kernel space) $\mathcal{S}$, i.e., $\phi : \mathcal{X} \rightarrow \mathcal{S}$. The mapping defines dot product $x_1^T x_2$ in the kernel space as $\phi^T(x_1)\phi(x_2)$. The direct mapping to space $\mathcal{S}$ is implicitly performed by selecting a feature space which supports the direct computation of dot product using a nonlinear function in input space. The kernel function k which performs such a mapping is defined as

$$K(x, x') = \langle \phi(x), \phi(x') \rangle = \phi^T(x)\phi(x')$$

The expression shows that mapping to space S by function K happens implicitly without considering the actual form of ϕ. The kernel space equivalent of the squared distance $\mathcal{D}^2(x_1, x_2)$ is defined as $K(x_1, x_1) + K(x_2, x_2) - 2K(x_1, x_2)$. Therefore; kernel space representation of equation (5.2.1) is expressed as

$$F^{\phi(x_1),\phi(x_2)}(\phi(x)) = \frac{K(x_1, x_1) - K(x_1, x) + K(x_2, x) - K(x_1, x_2)}{\sqrt{K(x_1, x_1) - 2K(x_1, x_2) + K(x_2, x_2)}} \tag{5.2.2}$$

The above expression gives the formulation of line projection in kernel space defined by pivots $(\phi(x_1), \phi(x_2))$. We define the set of intervals $[t_1, t_2]$ for binarization of equation (5.2.2) as

$$V(\phi(x_1), \phi(x_2)) = \{[t_1, t_2] | \Pr_{x \in \mathcal{X}}(F_{t_1,t_2}^{\phi(x_1),\phi(x_2)}(\phi(x)) = 0) = 0.5\} \tag{5.2.3}$$

The kernel distance based hash function family is defined as

$$\mathbb{H}_{\mathbb{KDBH}} = \{F_{t_1,t_2}^{\phi(x_1),\phi(x_2)}(\phi(x)) | x_1, x_2 \in \mathcal{X}, [t_1, t_2] \in V(\phi(x_1), \phi(x_2))\} \tag{5.2.4}$$

An indexing and retrieval scheme can be formulated using the hash function family $\mathbb{H}_{\mathbb{KDBH}}$ as discussed in section 4.4. The hash table parameters L and k have similar implication as in the indexing with traditional distance based hashing.

5.3 Multiple Kernel Learning for Hashing

Equation (5.2.2) represents the line projection function in kernel space and defines the mapping function for Kernel DBH. The optimum kernel selection is an important task in kernel learning based methods. The MKL methods use set of kernels instead of selecting one specific kernel function for learning the decision boundary in kernel space. The learning process of the algorithm selects the optimum kernel as the combination of base kernels. The algorithm removes the dependency of kernel methods over cross-validation for optimal kernel search. Additionally, the selection of specific kernel may induce bias in the solution. In this case, the MKL based combination of kernels over single kernel gives more robust solution. In practice many vision problems involve multiple input features; MKL provides an efficient way to combine them as different features have different similarity measures. In the following discussion, we present novel concept of MKL in indexing applications, so as to apply multiple features to improve the indexing performance. Our framework uses Kernel DBH function for generating indexing data structure. In this section, we discuss the limitation of existing MKL schemes in the context of our problem and present a new GA based optimization framework which overcomes the limitations of the classical approach for indexing applications.

5.3.1 Optimization Problem Formulation

The proposed framework selects the kernel K for hashing in the equation (5.2.2) by defining a learning based framework. The composite kernel K is defined as the parametrized linear combination of n base kernels, i.e., $K(x_1, x_2) = \sum_{i=1}^{n} w_i K_i(x_1, x_2)$ with $w_i \geq 0 \ \forall i$. The non-negativity constraint enforces Mercer's condition on kernel K. The proposed learning framework intends to learn optimum hashing kernel for indexing; therefore, we define retrieval performance maximization as the optimization objective. The nature of optimization objective requires the learning to be performed in semi-supervised setting, which utilizes subset of training examples available with label information during training. The ideal case of Distance based hashing should generate a hash table assigning unique index to all the examples belonging to a category. However, feature similarity based hashing does not guarantee unique hash index for similar category objects because of the semantic gap between low level features and high level semantics. The category information available with an example represent significant amount of inbuilt semantics. Therefore, the utilization of category information available with the training examples in a semi-supervised framework enforces more realistic grouping of objects in hash space.

The limited amount of labelled data may create the condition of over fitting. Therefore, the optimization objective requires a regularizer term to ensure desirable indexing performance independent of the amount of labelled training data. We apply the maximum entropy principle based regularizer presented by Wang et al. [321]. The objective of the regularizer is to maximize the information provided by each function value $h(\cdot)$ in ob-

ject index $g(\cdot)$. The regularizer is implemented by applying maximum entropy principle which assigns equal probability for function value $h(\cdot)$ to be 0 or 1; therefore, generating balanced partition of data in hash space. Following the result presented in [321], variance maximization of $h(X)$ satisfies the maximum entropy condition for function h. Therefore, the complete optimization objective for MKL problem is defined as:

$$\begin{aligned} \mathbf{w}^* &= \texttt{argmax}_{\mathbf{w}} \ \ \mathbf{F}(X, X_v, \mathbf{w}) \\ \mathbf{F} &= \mathbf{J}(X_v, \mathbf{w}) + \lambda \mathbf{V}(X, \mathbf{w}) \end{aligned} \tag{5.3.1}$$

X represents the complete training set, X_v represents the subset of training examples assumed to be available with label information for which function **F** is evaluated for different weight parameter **w**. λ represents the regularization parameter. Function **J**(X_v, **w**) represents the retrieval performance of KernelDBH computed over X_v and **w**.

$$\mathbf{J}(X_v, \mathbf{w}) = \texttt{mean} \left\{ \sum_{x_i \in X_v} \Delta(y_i, \hat{y}(x_i, \mathbf{w})) \right\}$$

$\hat{y}(x_i, \mathbf{w})$ represents the set of retrieved results for the validation query $x_i \in X_v$. y_i represents the actual label for the query and $\Delta(y_i, \hat{y}(x_i, \mathbf{w}))$ represents the computed retrieval score for x_i. Function **V**(X_v, **w**) represents regularizer term defined as the sum of variance of the hash values for all hash tables which is computed as

$$\mathbf{V}(X, \mathbf{w}) = \sum_{i=1}^{L} \sum_{j=1}^{k} \texttt{Variance}\{h_{ij}(X, \mathbf{w})\}$$

5.3.2 Genetic Algorithm based Optimization Framework for Multiple Kernel Learning

The kernel combination weights ($\mathbf{w}$) in equation (5.3.1) are the optimization parameter. The existing literature in MKL methods have various optimization problem formulations. The earliest MKL formulations defined the learning problem as Quadratically constrained quadratic program [167] and Semi infinite linear program [293]. Rakotomamonjy et al. [252] simplified the learning problem by moving to weighted 2-norm regularization formulation. In [111], the MKL formulations based on Boosting methods is presented. These MKL methods, primary developed for classification and recognition problems define joint optimization task for decision boundary and optimal kernel learning. The optimization objective of these MKL formulations are continuous in nature and follow the conventional gradient based methods for solution. However, the proposed MKL formulation for indexing evolves as an unique class of optimization objective (Equation (5.3.1)). The current optimization function is discrete in nature whereas the optimization parameter space is continuous. The non-differentiable nature of the optimization function restricts the application of existing gradient based solutions for proposed MKL.

For such optimization tasks, meta-heuristic optimization algorithms provide efficient solution for searching large parameter spaces. The Genetic algorithm is class of Evolutionary algorithms which is extremely suitable for global optimal parameter search in complex spaces [113]. GA follows the fundamental of natural evolution to explore the large and complex parameter spaces, and performs intelligent random search to quickly

identify the best solution. The primary advantage of GA based optimization comes from its parallel nature of parameter search. The GA process samples candidate solutions in different directions of parameter space and evaluates their suitability based on their fitness value. The iterative optimization proceeds the search direction towards the prominent candidate solutions; therefore, providing a robust and fast methodology to search through large parameter spaces. The multiple direction search also reduces the possibility of local optimas as the best solution. Considering the complex nature of parameter space, exhaustive search based solution for the current optimization problem is unacceptable. Therefore, we propose the MKL formulation for indexing in GA based optimization paradigm.

The parameter search is started by creating initial population as set of strings representing chromosomes. The population strings in GA represent the genetically encoded set of possible solutions. The search follows the evolutionary process defined by a set of genetic operations performed over population strings. The evolutionary process retains the potential strings and uses them for successive population generation. The selection of potential strings, i.e., candidate strings for successive regeneration, is based on their fitness values. The fitness value represents suitability of the population strings for the current problem. The fitness function in GA should be closely related to optimization objective. As maximization of retrieval performance defines the objective of learning problem, we define function **F** as the fitness function. The training data subset X_v available with label information is used as validation query set.

In the present GA optimization framework (Refer table 5.1), a population string represents the concatenation of genetic encoding of kernel weight parameters w_i. The

genetic encoding of optimization parameters allows the application of different genetic operators to proceed the search process. The fitness evaluation for each population string

Table 5.1: Algorithm: GA for MKL

1	Population generation ⇒ *Pop* = `Generate`()
2	Population initialization ⇒ `Initialization`(*Pop*): `Evaluate`(**F**) for each string in *Pop*
	For each **i** $<$ **noIterations**
3	Selection of individuals for successive population generation ⇒ *Pop_sub* = `Selection`(*Pop*) using *Tournament Selection*
4	Offspring generation step 1 ⇒ *Pop_1* = `Crossover`(*Pop*) using *Single Point Crossover*
5	Offspring generation step 2 ⇒ *Pop_2* = `Mutation`(*Pop_1*) using *Uniform Mutation*
6	Evaluate Offspring ⇒ `Evaluate`(**F**) for each string in *Pop_2*
7	New population generation ⇒ *Pop_new* = `Generate`(*Pop*, *Pop_2*) using *Elitist Selection*
	End

is done by measuring the retrieval performance of indexing using X_v as the query set. Initial population strings are randomly generated. Tournament selection based approach is used for selection of potential strings for successive generation. Tournament selection is easier to implement and works efficiently on parallel and no-parallel architectures. Additionally, it provides a simple way control its performance by optimizing tournament size. The tournament selection method selects p individuals randomly from the population, and the string with highest fitness in the selected p is placed in Mating pool. For a population vector having M strings, the process is repeated M times. The selected strings regarded as parents are subsequently used for generating the new population strings. Crossover and Mutation define to steps to generate new string for exploring the parameter space. We apply Single point crossover and Uniform mutation over the Mating pool for offspring

generation. The Single point crossover selects a crossover point, and swaps the part of parent strings defined by crossover point position. Crossover operation is controlled by crossover probability p_c which defines the percentage of individual strings from the mating pool to be used for generation. It is normally selected from $0.6 \sim 0.9$, i.e., using most of the mating strings for generation. The generation strings are subsequently applied to mutation operator. Mutation helps to search the unexplored search space by enhancing the variability of reconstructed strings by generating new string by single parent. Here, we have applied uniform mutation over the strings generated by crossover operation. Uniform mutation inverts each bit of parent string with probability p_m defined as mutation probability. It is normally selected between $0.02 \sim 0.05$. Higher values of p_m deviates the algorithm from optimal search path, therefore terming it as random search process. The set of offspring obtained by crossover and mutation are further used to generate new population. New population is constructed by Elitist selection based strategy using the old population and offspring set. We use Elitist selection based strategy for constructing new population as it preserves the better individuals in the current population for successive evolution process from both the old and new population.

5.3.3 Preliminary Evaluation with MNIST dataset

Initial evaluation of the proposed feature combination based indexing scheme is performed on MNIST dataset [173]. The experiment evaluates the Kernel DBH based indexing scheme by learning the optimal kernel by proposed MKL. MNIST dataset is a collection of handwritten digit images containing 60000 training and 10000 test images. Each dataset

example represents 28×28 grayscale image displaying, an isolated digit between 0 to 9. In this experiment, we have considered grayscale pixel intensities as features, i.e., each image is represented by a 784-d vector.

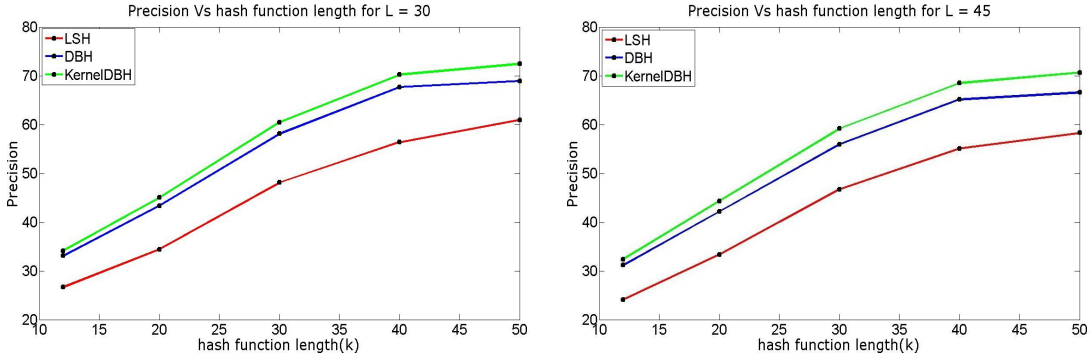
The initial population in GA optimization framework consisted 40 strings. The crossover and mutation probabilities are selected as 0.8 and 0.05, respectively. The kernel weight parameter (w_i) is encoded by a 5-bit binary string. The regularization parameter λ is set as 0.25 for all the experiments discussed in this work. After learning the kernel weight vector ($\mathbf{w}$), the indexing and retrieval process follows the conventional DBH based indexing scheme (Refer section 4.4). The parameter selection is an important issue in GA based optimization. GA approaches the best solution by adaptive operations on the population strings. The optimization convergence depends on the choice of adaptation strategy and parameters. Significant amount of research has gone into development and analysis of GA optimization, however, a general methodology for deciding GA parameters for any problem is unavailable [66, 61]. The basic GA parameters for experiments in this work have been selected following the recommendations in [61]. The random nature of GA requires sufficient number of function evaluations to arrive at the best solution. Here, total function evaluations are computed as product of the number of population strings to maximum generations. For a difficult function, more number of function evaluations are needed. The difficulty of the objective function is described in terms of multi-modality, deception, isolation and collateral noise in the search path [66]. The structure of the optimization objective (Equation (5.3.1)) is complex to describe, therefore the number of generations is finalized based on initial trials to ensure the convergence is achieved. The

experiment is simulated for maximum of 100, and 150 generations to establish the validity of GA based optimization.

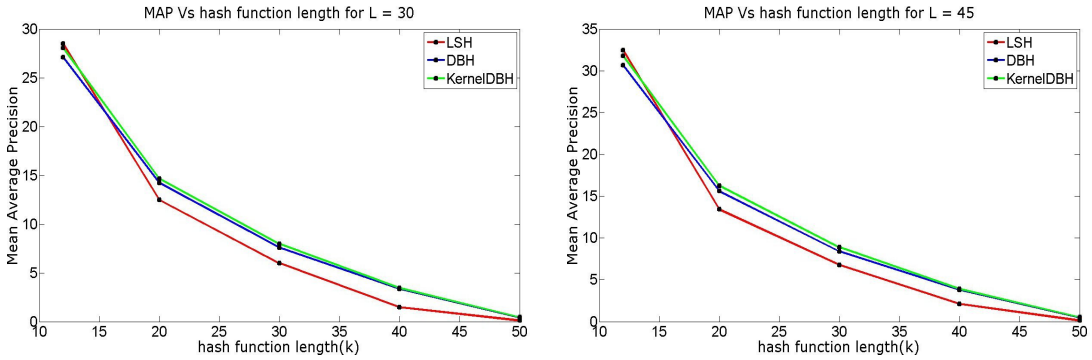
The base kernel set for learning the optimum kernel for hashing included set of 10 Gaussian kernels having variances from 100 to 900 on a linear scale. For learning the $\mathbb{H}_{\mathbb{KDBH}}$, the evaluation set X_v contains 1000 images selected by stratified sampling from the training images. The function generation is performed by stratified sampling of 1000 images from remaining training images (excluding the images selected in the evaluation set). The hash table generation is performed over 59000 training images. The ranking of retrieved results is an important performance measure for an indexing and retrieval scheme. The traditional retrieval measures, i.e., precision and recall are computed over unordered sets. However, Mean average precision (MAP) measure for query set X_v is computed over the ranked retrieval results [200]. The MAP computation procedure is explained in section 4.7.3.In the present indexing scheme, average precision for query q is computed over the collection of neighbours obtained from L hash tables having indices $g_i(q)$ for $i = 1, ..., L$. The neighbour ranking is performed based on the Euclidean distance from the query.

For performance measurement, the hash function families $\mathbb{H}_{\mathbb{DBH}}$ & $\mathbb{H}_{\mathbb{KDBH}}$, have been generated by stratified sampling of 1000 images from complete set of training images. The complete test image set has been used for performance evaluation. The LSH based indexing results for the similar hashing parameters are also presented. To define LSH based indexing, random hyperplane based hashing functions have been applied [44]. We have considered three parameters: Precision, MAP and Average number of comparisons for performance measurement which have been commonly used for the evaluation of

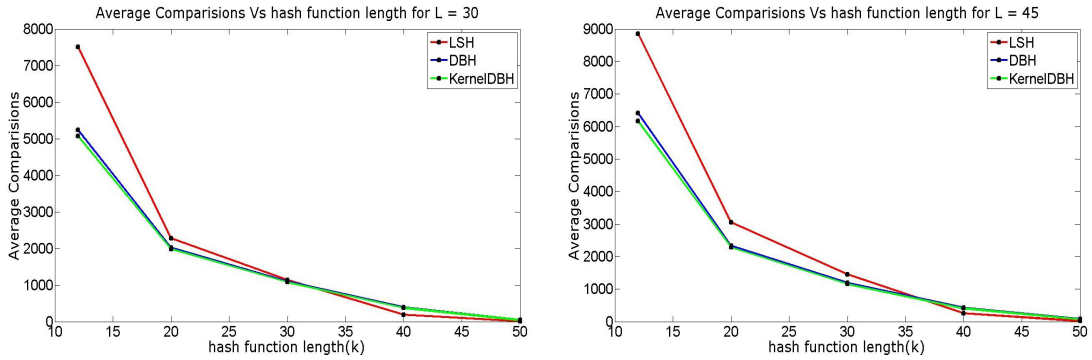
retrieval methods [320, 321, 220]. The number of comparisons represent the retrieval time complexity. Precision and Recall represent the grouping property of indexing scheme in hash space, with precision being the probability of the relevance of retrieved results and recall being the probability of the retrieval of relevant results. It is always desired to have high values of both measures; however, independent consideration neglects the ranking information of retrieved results. In this context, MAP presents single point measure of ranked retrieval performance by computing the average precision over the complete recall scale. Therefore, MAP is adopted as the primary performance measure. The KernelDBH shows significant improvement in precision values over LSH and DBH with less number of comparisons. The experimental results for different hashing parameters and number of generations are presented in figure 5.2 and 5.3. The variations in performance measures for selected number of generations exhibit similar trend. However, $0.5 \sim 7\%$ of relative variation is observed in simulating for different number of generations because of the random nature of GA. Clearly, 100 generations are sufficient for GA to identify the best solution for the problem. For all the hashing methods, we observe sharp decrease in MAP score with increase in hash function length as the collision probability decreases exponentially with increase in k, resulting sharp decrease in recall. The KernelDBH and DBH achieved better precision and MAP values with LSH requiring less number of average computations for larger k. It is justified as the random hyperplane based projections are independent of data distribution; therefore, for smaller k, the hashing function g shows poor discriminative power. However, in practice, short hash functions are preferred at acceptable retrieval performance to control the size of indexing data structure. The experimental



(a) Precision Values



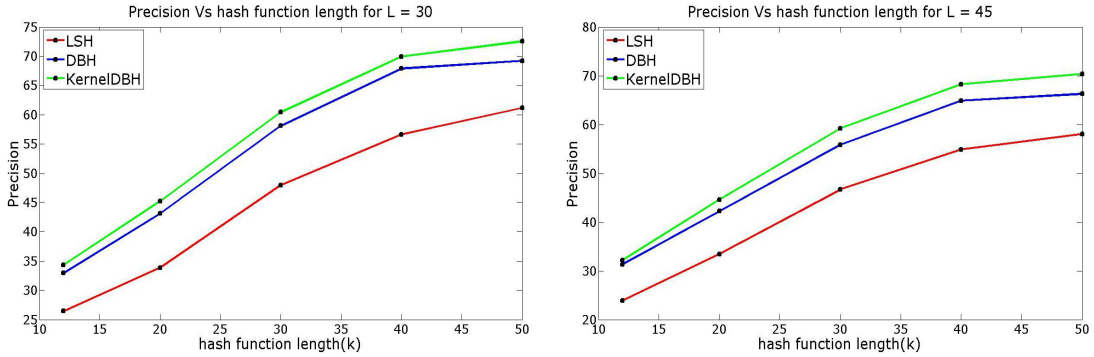
(b) Mean Average Precisions



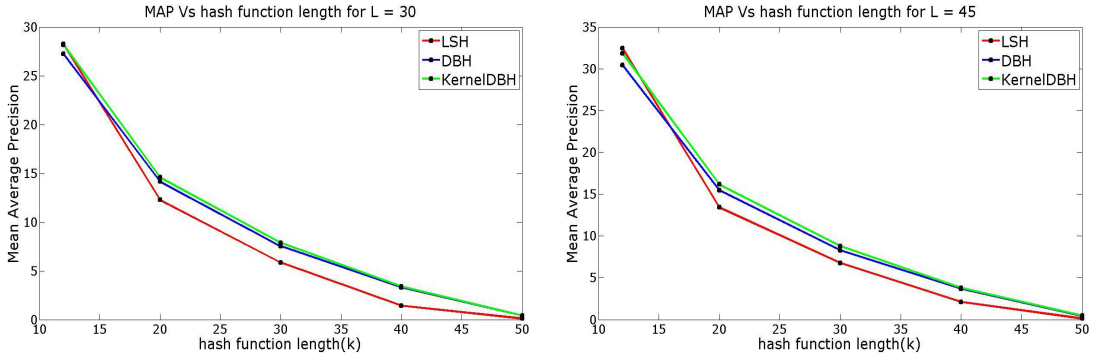
(c) Average Comparisons

Figure 5.2: Results with MNIST dataset: 100 generations

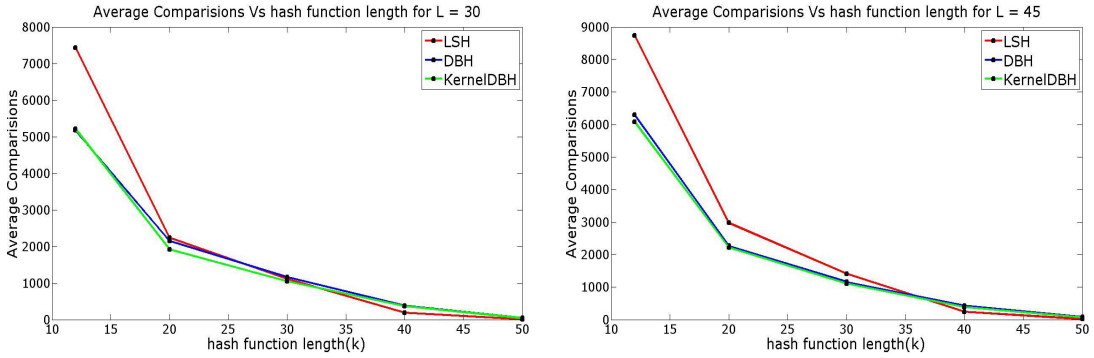
results validate the effectiveness of the proposed MKL framework for indexing. The results also demonstrate excellent grouping capability of the Kernel based DBH shown by



(a) Precision Values



(b) Mean Average Precisions



(c) Average Comparisons

Figure 5.3: Results with MNIST dataset: 150 generations

significant reduction in average comparisons for approximate nearest neighbour search.

The proposed concept presents a novel indexing scheme based on MKL based feature

combination. Nevertheless, a small set of experiments is performed to evaluate the effectiveness of proposed concept for recognition problem. The approximate nearest neighbour (NN) has been extensively applied for recognition based applications. The objective is to apply the approximate NN search for classification by grouping the examples in learned kernel distance based hashing space. The base kernels and basic GA parameters are fixed as in the previous experiment. Figure 5.2 and 5.3 show that maximum of 100 generations are sufficient for GA to converge to the best solution. In this case, the objective of MKL (Equation (4.7.1)) is defined as the maximization of classification accuracy. The Euclidean distance is applied for similarity measurement. The classification accuracy is computed for majority voting in 3 nearest neighbours. The results are presented in the table 5.2.

Table 5.2: Classification accuracies using the proposed scheme

		LSH results		DBH results		KernelDBH results	
Hashing paras.		Acc.	Avg. Comps.	Acc.	Avg. Comps.	Acc.	Avg. Comps.
$L = 60$	$k = 24$	94.50	7648	91.38	5276	92.73	5118
	$k = 40$	86.43	2375	95.65	2143	97.12	2086
	$k = 60$	67.37	1191	81.44	1156	84.65	1144
$L = 90$	$k = 24$	95.95	8823	93.27	6412	94.79	6124
	$k = 40$	88.04	3026	96.41	2576	97.86	2436
	$k = 60$	69.73	1546	82.57	1252	86.27	1228
1-Vs-1 SVM with Gaussian kernel [19]						98.57	

We compare the proposed concept with SVM which is widely accepted state-of-the-art classifier. The multi-class classification is performed by arranging pair-wise binary SVMs in the direct acyclic graph architecture. In this case, the Gaussian kernel based SVMs achieved classification accuracy of 98.57% with average 5451 kernel computations

[19]. The memory requirement needed to store 3175 support vectors. The KernelDBH based classification achieved the best accuracy of 97.86% with approximately 45% less distance computations. Also the table 5.2 shows that LSH based classification performs better for smaller hashing functions but requires more computations. For similar size of indexing data structure (size of indexing data structure is proportional to L and k), DBH and KernelDBH are more accurate and require less computations. In general, time complexity of SVM based classification in large class problems is much higher than approximate NN based methods. Additionally, hashing based methods provide the flexibility to tune the performance based on the real-time requirements, which is not available in SVM based classification. The results in table 5.2 show that proposed concept is an efficient method for approximate NN based classification. The SVM based classification does not provide any information about the similar objects whereas approximate NN retrieves set of similar objects at lesser computational cost. However, the efficacy of the method in comparison with recent classification algorithms requires deeper experimental and theoretical analysis.

The above experiments establish the ability of presented concept in learning the optimal kernel for the classification. However, the image indexing and retrieval is the major application considered in this work. In the case of base kernels computed from different features, the concept provides a more principled and logical approach for applying multiple features in developing the indexing and retrieval applications. In the next section, the experimental validation of the statement is presented by applying the proposed concept for developing two practical indexing and retrieval applications using multiple feature representations.

5.4 Document Image Indexing Using Combinations of Features

Word based document indexing provide efficient solution for managing old and degraded documents, and for the scripts for which reliable optical character recognizers (OCRs) are not available. The indexing scheme here exploits the image properties of the document. The textual contents in the document i.e. word images are used for generating indices. As discussed in chapter 4, the problem poses two primary challenges in terms of unique feature representation for word images, and indexing computationally simple method for indexing. Chapter 4 applied single feature representation for document image indexing. Here, we explore the learning based combination of multiple features for document image indexing. Various script dependent/independent frameworks have been proposed in this direction [199, 205, 269, 190, 193, 20]. These works have presented novel word image representations and frameworks to support fast document retrieval. However development of word image based document indexing applications for Indian scripts have not been much researched. Additionally, the OCR technology for Indian scripts is still in developing stage [119]. Therefore; major evaluation of the scheme is done for indexing the documents of Indian scripts.

We select two document image collections belonging to Devanagari and Bengali script for evaluation. Figure 5.4 shows the proposed document image indexing scheme using the Kernel based DBH. Once the kernel weight vector $\mathbf{w}$ is learned by the MKL process, the offline process of indexing generates $\mathbb{H}_{\mathbb{KDBH}}$ and hash tables containing word

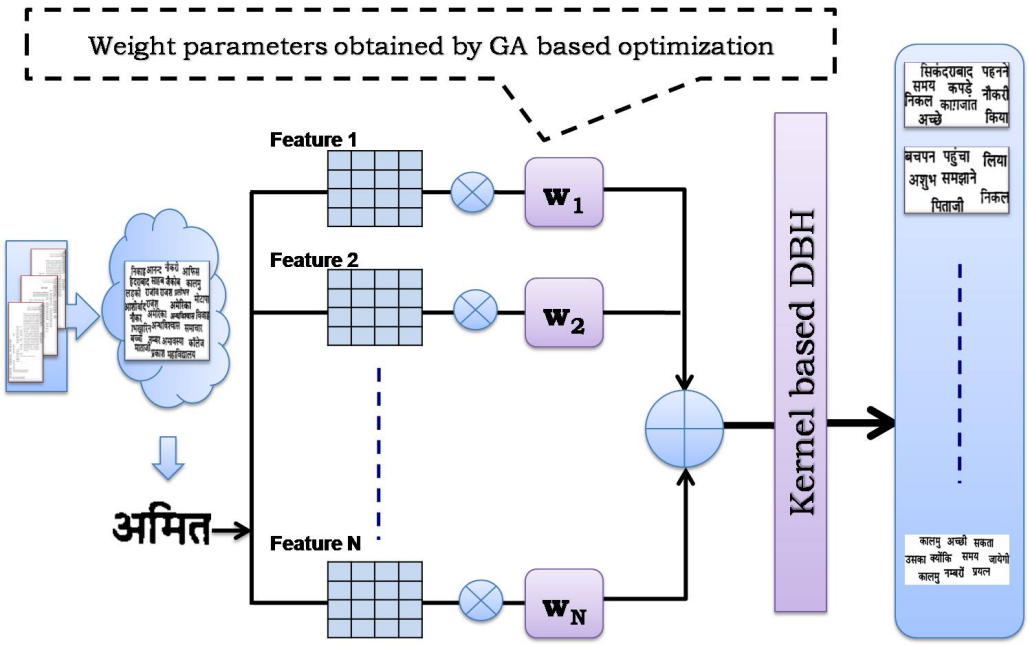


Figure 5.4: Kernel DBH based document image indexing

indices. The retrieval is performed as online processing following the steps as discussed in section 4.2. We have proposed a novel set of features computed using the shape properties of word images. The features individually represent different level of global and local shape information. The proposed MKL process learns the optimal combination of these features for indexing to improve the performance.

The experiment is performed on sampled Devanagari and Bengali script document image collection available at [15]. The Devanagari document collection consists of 503 pages scanned from 6 books. The Bengali document collection consists of 226 pages scanned from 4 books. The collection is prepared by scanning old Indian script books at the resolution of 300dpi. The scanned images are of comparatively low quality, primarily because of the degradation in original document pages. The preprocessing step included

smoothing and de-skewing. The binarization step is performed by Otsu's method [232]. The word segmentation is performed by horizontal and vertical profile based technique. The selection of meaningful word images for generating document indices is done by applying word length in terms of pixels as criterion. Additionally, the stop words, punctuation and typographical marks are filtered out by applying conventional thresholds (aspect-ratio, word length). After initial filtering, Devanagari word collection consists of 23145 words, and Bengali word collection consists of 18632 words. The feature extraction details and experimental results of the proposed indexing scheme are presented as following.

5.4.1 Feature Description

Shape information represents important visual cue for object recognition problems. For this work, the descriptor computation steps discussed in section 4.3 have been applied for defining two feature representations by following different point extraction methods.

The first methodology for descriptor point extraction involves envelope curve detection of word image (Figure 5.5). The method for extraction of envelope curves is discussed

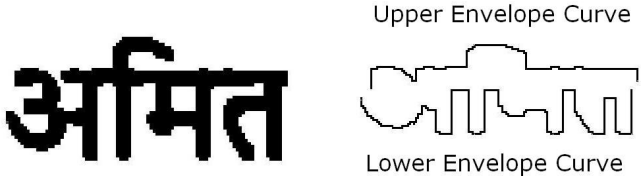


Figure 5.5: Envelope curve of the word

in [21]. The envelope curves of a word are considerably tolerant to different font properties and incorporate font invariance in the descriptor. The point set P obtained by uniform

sampling over the envelope curves define the set of descriptor points. The point set P is further used to compute the Envelope Curve based Shape Descriptor (ECBSD), which represent the outer boundary information of word shape. The descriptor computation follows the steps discussed in section 4.3 (Refer the section for details). We represent the computational steps for clarity.

- Sampling of point set P from the envelope curve of the word object.
- Shape context computation for all points in P.
- Partition-wise pdh computation with respect to points located in the partition
- $h_{final} = \{h_1 : h_2 : ... : h_{\text{num\_parts}}\}$, here h_i represents pdh for i^{th} partition
- Fourier transform of h_{final} defines the shape descriptor $F(P)$

In many Indian scripts, some characters have outer boundary similarity, e.g., {/sha, /pa}, {/ba, /va} character pairs in Devanagari and {/k, /ph}, {/dz, /sh, /dh} and {/dh, /ch} character pairs in Bengali (Figure 5.6). In this case, the appearance of inner contours distinguishes their semantics. The shape descriptor (SD) presented in section 4.3 provides

Figure 5.6: Example character images

unique representation to such characters by following grid based approach for descriptor point extraction. The process helps extraction of shape information inherent in the inner

contours of word image for defining their unique representation. We use shape descriptor as the other feature representation for document indexing experiment. Clearly, ECBSD concentrates on the global characteristics whereas SD concentrates on local characteristics of word shape. The optimal combination of these complimentary nature of informations can significantly improve the indexing performance by refining the word groups generated by hashing.

5.4.2 Retrieval Results

The preprocessing steps for ECBSD and SD computation include bound detection and size normalization by aspect-ratio preserving scaling transform. The selection of number of partitions is based on heuristic knowledge as most of the Devanagari and Bengali scripts words are formulated by combination of 3 to 5 characters. Therefore, for both the scripts, we divide the word image in 4 partitions for descriptor computation. For SD computation, the placement of logical grid is done at interval of 4 pixels horizontally and vertically. For Devanagari word images, the pdh dimension for each partition is selected as 40×30, i.e., shape context for each point is computed for 40 distance and 30 angle bins. For Bengali word images, the pdh dimension for each partition is selected as 40×36. For both the word image collection, similar descriptor parameters (number of partitions, pdh dimension) have been selected for ECBSD and SD computation. The frequency of occurrence for different words in the document collection contains large variation. In this case, the random selection and heuristic based methods for pivot object selection do not exploit the complete distance distribution information. We follow clustering based

approach for pivot object selection. For defining the indexing scheme for a document collection, feature based clustering is performed over complete word image collection. The cluster centres are selected pivot object set. Cluster center computation is done as the mean of objects belonging to the cluster. The selection of mean as cluster center increases the robustness of indexing scheme by reducing the effect of outliers. The selection of number of clusters requires the prior information about the frequency of different words. Therefore, we apply density-based clustering algorithm proposed in [86] which estimates the number of clusters by density estimation for the given input parameters.

We applied MKL to learn the optimum kernel for indexing using both the descriptors. For Devanagari script collection, the base kernel set for both descriptors consisted of linear, and Gaussian kernels with variance = {1, 4, 10, 20, 100}. For Bengali script collection, the base kernel set for both descriptors consisted of linear, and Gaussian kernels with variance = {1, 2, 5, 20, 40, 100}. For GA fitness evaluation, MAP measure over the retrieval results for validation query set X_v is computed. The validation query set for Devanagari and Bengali document indexing consists of 217 word images, and 179 images respectively. The GA population consists of 40 individuals. The kernel weight parameter is encoded by 5-bit binary string. The crossover and mutation probabilities are selected as 0.6, and 0.05 respectively. The GA iteration is terminated after simulating 200 generations. These parameters and criterion have been constant through all the document indexing experiments. For learning the optimal combination of descriptors for indexing, the base kernel set is formulated by ORing the individual base kernel sets corresponding to each descriptor. The retrieval result using the DBH based document indexing with similar

hashing parameters is presented as baseline. The final evaluation of Devanagari script, and Bengali script document indexing is done for 177 and 209 query images respectively. The retrieval results for different hash table parameters are presented in table 5.3 and 5.4. The indexing performance by SD and ECBSD features achieve comparable performance. The MAP scores achieved by KernelDBH based indexing have significant improvement $(3.54 \sim 8.47\%)$ over the baseline DBH based indexing performances. Additionally, objects grouping in hashing space is also improved by KernelDBH thereby reducing the average number of comparisons $(2.11 \sim 7.31\%)$ over baseline DBH results. In addition to dominating global and local shape characteristics in ECBSD and SD respectively, both the descriptors contain fair amount of overlapping information. Combining both the

Table 5.3: Retrieval results with Devanagari script

	Hash table parameter L = 40					
	k = 12		k = 20		k = 30	
	MAP	Avg Comp	MAP	Avg Comp	MAP	Avg Comp
DBH(SD)	74.51	1703	71.04	1064	65.86	722
DBH(ECBSD)	81.11	1522	74.68	823	66.39	412
KernelDBH(SD)	78.15	1589	74.98	1013	70.13	669
KernelDBH(ECBSD)	84.08	1476	78.33	813	70.87	398
KernelDBH(SD+ECBSD)	**85.95**	1435	**81.93**	804	**75.30**	384
	Hash table parameter L = 50					
	k = 12		k = 20		k = 30	
	MAP	Avg Comp	MAP	Avg Comp	MAP	Avg Comp
DBH(SD)	75.90	1769	71.45	1235	66.20	777
DBH(ECBSD)	82.21	1630	75.31	902	67.71	445
KernelDBH(SD)	79.68	1644	78.41	1178	70.44	721
KernelDBH(ECBSD)	85.34	1588	79.64	864	72.08	432
KernelDBH(SD+ECBSD)	**86.87**	1536	**82.39**	844	**76.18**	417

Table 5.4: Retrieval results with Bengali script

	Hash table parameter L = 40					
	k = 12		k = 20		k = 30	
	MAP	Avg Comp	MAP	Avg Comp	MAP	Avg Comp
DBH(SD)	77.26	1458	70.55	857	61.99	507
DBH(ECBSD)	83.53	1449	74.59	759	62.66	285
KernelDBH(SD)	80.16	1411	73.62	787	65.73	498
KernelDBH(ECBSD)	86.26	1421	77.96	712	67.75	281
KernelDBH(SD+ECBSD)	**88.11**	1395	**81.35**	685	**71.24**	272
	Hash table parameter L = 50					
	k = 12		k = 20		k = 30	
	MAP	Avg Comp	MAP	Avg Comp	MAP	Avg Comp
DBH(SD)	78.38	1474	71.00	892	62.11	519
DBH(ECBSD)	84.99	1485	75.42	781	63.80	308
KernelDBH(SD)	80.65	1428	73.86	854	66.27	508
KernelDBH(ECBSD)	87.03	1446	78.78	737	67.83	305
KernelDBH(SD+ECBSD)	**88.90**	1417	**82.14**	715	**72.08**	299

descriptors in kernel space, the MKL framework carefully selects the distinct attributes to increase the discriminating power of the resulting descriptor. The improvement in the indexing performance $\{(4.82 \sim 11.51\%)$ increase in MAP scores and $(2.37 \sim 8.76\%)$ reduction in average number of comparisons over baseline DBH based indexing$\}$ by MKL based combination of features establish the effectiveness of the proposed scheme. For the Devanagari image collection, the text version prepared by recognizer discussed in [12] achieved MAP score of 81.24% for same query set. The proposed scheme here shows best MAP score of 86.87%. The result establishes the effectiveness of proposed indexing scheme when mature OCR technology is not available. Additionally, the proposed scheme has advantage of adaptive parameters which are adjustable as per real-time requirements.

The image based indexing is specifically useful for managing the documents for which reliable OCR's are not available. Therefore, the applicability of the proposed indexing scheme is demonstrated on Indian script document collection. Additionally, the applicability of the proposed scheme is also validated for indexing English script document collection. The objective is performance comparison of proposed indexing and retrieval scheme with the retrieval in OCR'ed form of document images. The GA based optimization parameters are same as the earlier experiment in this section. For the purpose of benchmarking, text version of the collection is created by commercial OCR available at [303]. The experimental collection comprises of 212 pages from 6 books compiled by sampling document images from the Google book dataset [116]. The images have been used in original form without any preprocessing. The conversion of original gray scale images to binary images is performed by Otsu's method. The word segmentation from the document images is done by horizontal and vertical profile based technique. After initial filtering, the word image set consisted of 19721 words. The filtering process removes stop words, punctuation marks, and words having length less than preselected threshold. The word image representation is generated by features discussed in section 5.4.1. The majority of English words are formed by 4 to 7 alphabets; therefore, shape descriptor is computed using 1×6 partitions. The pdh dimension for each partition is selected as 38×36. The strategy for index structure creation is followed as discussed earlier. The base kernel set for both descriptors consisted of linear, and Gaussian kernels with variance = {1, 2, 5, 25, 40 , 60, 100} and MAP measure is considered for evaluating the GA fitness function. The query set X_v consisted 165 word images and performance

evaluation is done for 301 queries. The query word length varied from 4 to 11 characters. For learning the optimal combination of descriptors for indexing, the base kernel set is formulated by ORing the individual base kernel sets corresponding to each descriptor. The results presented in table 5.5 show best MAP measure of 91.68%. The Matlab based

Table 5.5: Retrieval results with English script

	Hash table parameter L = 40				Hash table parameter L = 50			
	k = 12		k = 20		k = 12		k = 20	
	MAP	Comps.	MAP	Comps.	MAP	Comps.	MAP	Comps.
DBH(SD)	78.64	1847	71.48	926	79.37	2157	72.15	1025
DBH(ECBSD)	85.32	1758	77.23	835	86.24	1930	77.67	964
KernelDBH(SD)	82.75	1734	76.31	888	83.06	2085	77.82	992
KernelDBH(ECBSD)	89.85	1638	81.58	801	89.95	1827	82.48	924
KernelDBH(SD + ECBSD)	**90.84**	1552	**84.87**	764	**91.68**	1736	**85.78**	894

implementation of the proposed scheme requires average 0.60 second for performing each query retrieval task. For recognizing the complete collection processing time of 189 second is consumed on 2.93GHz, 8GB RAM desktop computer. For same query set, retrieval over text version of document images achieved MAP score of 93.29%. Figure 5.7 shows the sampled images and recognized text. The boxes represent the retrieved words for query word 'Gaelic' by the KernelDBH based indexing. Correspondingly, the incorrect recognitions of the queried word and other wrong recognitions are encircled in the text. The primary advantage of partition based approach for feature computation is local handling of word shape degradations which imparts invariance to minor distortions in the descriptor. Additionally, the approach helps the retrieval based on parts-of-word similarity as shown in figure 5.8. The retrieval performance of image based indexing is marginally

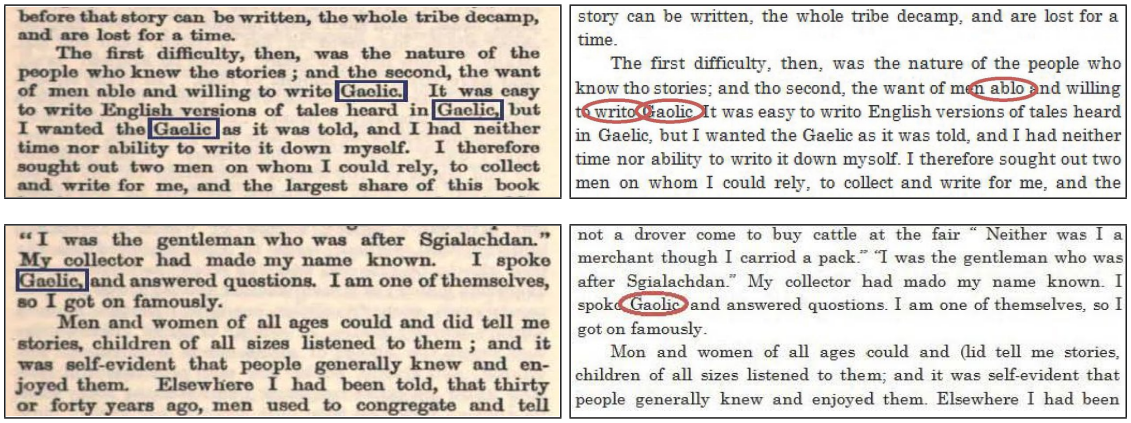


Figure 5.7: Sample images and corresponding recognized text placed side by side

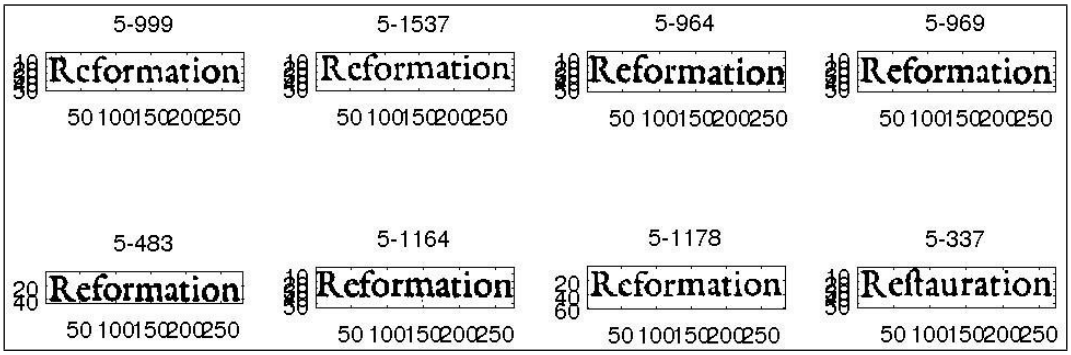


Figure 5.8: Retrieved words for query 'Reformation'

inferior in comparison with commercial OCR based retrieval. The OCR technology for Latin scripts have reached the state-of-the-art status but similar recognition performance for Indian script document images is not guaranteed because of under-developing OCR technology. Nevertheless, the proposed scheme is an efficient alternative for indexing such document collections which is shown by retrieval comparison for Devanagari script.

5.5 Conclusions

A novel image indexing and retrieval scheme which can efficiently utilize diverse and complementary informations represented by different feature extraction methods to improve the indexing performance. We have proposed novel extension of DBH to kernel space which provides the platform for MKL formulation in feature based indexing. The novel formulation provides an efficient and robust approach for application of multiple features in indexing problems by learning their optimal combination in kernel space. The proposed indexing scheme demonstrated novel application of Genetic algorithm for optimization task in MKL. The significant improvement in retrieval performance shown by experiments on different class of image collections (handwritten digits and document images) have validated the effectiveness and generalizability of proposed concepts. The learning based feature combination for indexing provides a principled approach for fusion of multiple modalities existing in documents. In the next chapter, we explore this aspect of the framework.

Chapter 6

Multi-modal Information Integration for Document Retrieval

6.1 Introduction

A document image is a presentation form which includes information in multiple modalities including graphics and images, other than textual component in the image format. This is equally valid for documents containing text in electronic form itself like .docx, .pdf files etc. We need to combine information from multiple media contents of documents for developing effective retrieval system. The previous chapters have looked into techniques for designing a retrieval system by using only image of the textual component. In this chapter, we address with the problem of integration of additional information embedded in the image component of documents as well as additional attributes of the textual component.

In particular, we consider documents having graphics/image and multi-script text

segments. Two important steps are required for the development of multi-script and multi-modal documents. First, the information of text and graphics segments and second the script identification of text segments. The chapter presents novel methods for identifying the text/graphics segments in documents and script identification of text segments. Once scripts are identified, text based retrieval engine can be built by applying an OCR to the text segment. However, degraded image quality and unavailability of mature recognition technology may restrict retrieval performance for OCR'ed output. We have proposed a new technique for dealing with the noisy text produced by OCR. In summary, this chapter describes following contributions:

- A scheme for segmentation of documents with complex layouts is proposed. The bottom-up framework for segmentation combines the clustering and conditional random fields (CRFs) based modelling. The approach assigns image pixels to dominant color modes representing different colour planes. The CRF based modelling is applied for final labelling which extracts the local neighbourhood information across different colour planes by learning the semantic correlation in color and image space. The experimental evaluation is performed on multi-coloured documents having overlapped text/graphics images.
- Multi-modal retrieval framework for document images is proposed. The framework uses the knowledge of text and graphics regions, and applies learning based feature combination for indexing discussed in the section 5.3 for multi-modal retrieval of document images.

- A new technique for script identification in documents is presented which addresses the general scenario of bi-lingual documents having mixed text of different scripts even at the level of sentences or words. The proposed scheme works in hierarchical fashion by identifying the script at block/paragraph level followed by identification at word level. First, the texture property of text segments is exploited for script identification by Relevance vector machine based classifier. Subsequently, we apply a rejection based classifier based on Adaboost for word level script identification by exploiting shape characteristics. The experimental validation is shown on document collections of mixed Hindi/English and Bengali/English texts.
- Indexing framework for text documents is proposed which addresses the issues of optical character recognition (OCR) errors in indexing process for improving the overall retrieval accuracy. The framework applies latent Dirichlet allocation (LDA) for generating the document indices. The recognizer's confidence characteristics in correctly recognizing a symbol is propagated to topic learning process such that semantic grouping of words carefully distinguishes between commonly confusing words. A novel application of *Lucene* with topic modelling is presented for document indexing application. The experimental evaluation is shown on document collection of Devanagari script.
- To improve retrieval performance, we have proposed a new word based document image retrieval technique by combining output of OCR and image based features. This technique uses the learning based feature combination for indexing discussed in section 5.3. We demonstrate that, retrieval performance on noisy text documents can

be improved by unified indexing with image based features. The topic distribution based representation for text, and shape based image feature is applied for document representation.

We start with the survey of existing work related to the problem domain. Subsequently, our contributions are presented.

6.2 Methods for Multi-modal Document Image Retrieval

6.2.1 Existing Text/Graphics Segmentation Methods for Document Analysis

Over the years several document layout analysis algorithms have been proposed in [41, 201]. These approaches are broadly divided in two categories namely top-down and bottom-up methods. Top-down approaches use global properties such as white space separations in the document page thereby partitioning the document image into component blocks. Such techniques fail while handling documents with complex layout as they lack clear separations. Conversely, Bottom-up approaches start at pixel level and propagate the local information at global level to perform segmentation. The method proposed in [165] uses a combination of top-down and bottom-up methodology for handling documents with complex layouts. Lin *et al.* [181] present a hybrid approach to segment and classify document content as text, picture and background. Mukherjee *et al.* [222] discuss multi-scale clustering based technique for document segmentation. Also these approaches are based

on classical document segmentation methodologies that are only applicable for gray-scale or binary images. Many model-guided segmentation and layout analysis schemes [48, 325] are also reported in literature. Mighlani *et al.* [215] perform color histogram analysis for automated layout segmentation of documents. Layout analysis using color information have been proposed in [334, 295, 26] to handle color document images with complex layouts such as forms, text overlaid on image, posters etc. These approaches work on either RGB distribution or use some color reduction algorithms to use an optimal set of color for text extraction only. [57] describes a methodology for extracting text rendered in uniform color. The algorithm uses connected component analysis based on color similarity in the RGB color space. Various researchers [161, 2] have reported usage of wavelet based techniques for document image understanding (extracting text, picture and background). However, such schemes require sufficient prior knowledge of similar documents and are computationally intensive. The document layout analysis is closely related with scene text detection in natural images primarily due to classification type of problems. However, text detection is basically binary classification problem, whereas document layout analysis is multi-class problem. In this context, some works in text detection in scene images are presented in [87, 322, 155, 174]. Early work presented in [338] developed texture based segmentation framework for text detection in images. Lienhart and Wernicke [180] introduced application of multi-layer feed-forward neural network for text line detection in images. In [82], stroke width transform is introduced which replicates scene image having pixel values as the width of the stroke having maximum probability of belongingness. Subsequently, these pixels are grouped into candidate letters based

on modified connected component analysis. Coates et al. [58] applied k-means based feature learning for character recognition. Recent work by Neumann and Matas [225] presented multi-stage paradigm for text detection in scene images. First stage is extremal region detector for different color channels based on set of local descriptors. Sequentially, distinct extremal regions are selected by a sequential classifier consisting of two stages having AdaBoost with decision tree, and Gaussian kernel SVM. The existing document image segmentation methods have not addressed scenario of overlapping text/graphics. We propose text/graphics segmentation methods for documents having complex layout. The method works through self-organizing decomposition process in color space. The approach extracts different logical components despite being overlapped in image space. Subsequently, CRF is applied to model the contextual dependencies in image space by using a new formulation of neighbourhood across clusters in color space and spatial proximity.

6.2.2 Existing Script Identification Methods

The automated script detection in document analysis performs the script-wise segmentation of text regions. Existing works in this direction have addressed this problem at different levels namely page, paragraph and word. The earlier works in this area, Spitz [294] examined upward concavities of connected components for distinguishing between Asian and European languages. Tan [301] used multiple channel Gabor filter based rotation invariant features for automatic script identification on Latin and south-asian script documents. Rashid *et al.* [274] present a discriminative learning approach for Greek -

Latin script identification from ancient document images at connected component level using convolutional neural network. Wood *et al.* [335] propose a method that exploited global characteristics of the text using Hough transform, morphological filtering and analysis of projection profile. Busch *et al.*[38] suggested use of wavelet based texture features, Gabor filter and gray level co-occurrence matrix for script identification. Hochberg *et al.* [134] describe an approach for automatic script identification using cluster based templates of textual symbols. The experimental evaluation have been presented on thirteen scripts including Devanagari.

In the context of Indian scripts, [50] applied Gabor energy features extracted from connected components in text segments and used them for recognizing Devanagari, English, Telugu and Malayalam scripts. In [47], authors demonstrated the combination of Gabor filter based techniques and direction distance histogram classifier for script identification. Basavaraj and Subbareddy [242] proposed neural network based system for identification of English, Devanagari and Kannada scripts. The authors used morphological operation as preprocessing and used direction based pixel distribution for feature representation. The approach is further extended in [69] for addressing the font variations. Sharma *et al.* [273] applied curvature features with Gaussian Mixture Model for document level script recognition considering eight major Indian scripts. However, the script variation at line and word level is common in multilingual environment. Sufficient amount of work is available on Indian script identification at line and word level. Amongst the earlier works are Pal and Chaudhuri [237, 236]. In [237], a tree based approach is presented for separating Roman, Devanagari and Bengali script words. The approach identifies the

Roman script based on the existence of headline features. Subsequently projection profile and stroke based features are applied for final classification. The extension of the approach for larger number of scripts is presented in [236]. However, the approach required prior information about the script triplet in documents. The work in [238], proposed identification of Latin, Chinese, Arabic, Devanagari and Bengali text lines. The headline and principal stroke features are used for separating Bengali and Devanagari from other scripts. Chinese is identified by checking the existence of characters with four or more vertical runs. Latin is separated from Arabic by using statistical and morphological features. Pal *et al.* [240] presented a generalized approach for identifying script without any prior knowledge in contrast to approach described in [236]. The existence of headline, maximum distance between consecutive characters, border pixels along with profiles and morphological features were applied for feature representation. Padma *et al.* [233] developed a model to identify script from a trilingual document printed in Kannada, Hindi and English using top and bottom profiles features.

Additionally, significant number of works have addressed script identification problem at word level. At word level, the identification completely depends on the type of features employed as information available from few characters in the word may not be sufficient. Features such as connected components, moments, compactness of shape description, stroke geometry etc are fragile to noise. Hence, script recognition at word level is a difficult task. In [145], author combine headline feature with contextual information to distinguish between Devanagari and Telugu script words for a bi-lingual OCR application. The schemes presented in [238] and [240] were extended in [285] by includ-

ing additional features such as shift below headline, deviation feature, loop, tick feature and left inclination feature. Five script pairs namely Devanagari/Bengali, Bengali/Latin, Malayalam/Latin, Gujrati/Latin and Telugu/Latin were considered. Following the above discussed approach, English, Urdu and Devanagari text identification is performed in [42] and English and Telugu script recognition is presented in [43]. Dhandra *et al.* [70] suggested a morphological based script identification technique. Eccentricity, aspect ratio, directional stroke features and average pixel distribution based features were used for separating Kannada, Devanagari and English words. Padma *et al.* [234] used discriminating features for identifying and separating Kannada, Hindi and English words. In [71], two approaches to identify script at word level in a bilingual document containing Roman and Tamil is presented. The first method uses spatial spread of a word together with character density. The second method analyses the directional energy distribution of a word using Gabor filter at suitable frequencies and orientations. Pati et. al [241] reported a word level script identification technique in a multi-script scenario for 11 Indian scripts. They used a combination of Gabor and discrete cosine transform (DCT) features. The existing methods for script identification at page, paragraph and word level do not provide computationally efficient solution for practical scenario of multi-lingual documents having random use of different script. We propose a hierarchical framework for script identification in documents having mixed use of script.

6.2.3 Methods Addressing the Recognition Inaccuracies for Document Retrieval

Topic model based indexing has been preferred approach for text based retrieval [135, 31]. These models have been extensively applied for document summarisation, and indexing applications [5, 318, 300, 351, 253]. The text documents in this case are converted to vector space model defined as Term-frequency by Inverse-document-frequency representation (*tf-idf*). Here, the text words define the terms. Topic based stochastic modelling explores the latent topical structure of documents by grouping semantically related terms to a topic. Therefore, topic based document representation presents an intuitive approach for defining semantic retrieval schemes. Nevertheless, topic based indexing and retrieval applications has not been explored much for document image collection. In this direction, some of the existing works have developed topic model based error correction framework as post-processing step of OCR [93, 331]. However, the error correction does not guarantee perfect OCR'ed output. The effect of recognition errors on retrieval using *tf-idf* based representation has been empirically studied by Taghva *et al.* [299]. Also, Walker *et al.* [316] have recently evaluated the robustness of different topic models in the case of erroneous OCR'ed documents. In this direction, our work presents a novel framework to enhance the robustness of topic model based indexing of OCR'ed text without requiring explicit error correction after digitization. The approach exploits performance characteristics of the recognizer to improve invariance of semantic grouping of documents with respect to recognition errors. In particular, this technique has been used for document

images of Devanagari for which OCR with limited accuracy is available. Additionally, the work presents novel application of LDA as very few attempts have been made in the past for its use in retrieving document images of Indian scripts.

6.3 Separation Framework for Multi-coloured Text/Graphics

Document image segmentation is a crucial pre-processing step for content extraction. Document images typically consists of text, graphics/half-tones and background. Several documents such as magazines and brochures contain very complex layout. Segmentation and layout understanding of such documents presents a challenging task because of the following reasons:

i Random placement of figures and text.

ii Complex (textured/colored) backgrounds.

iii Text overlaid on images/graphical patterns.

iv Variation in text formatting in terms of font properties (font type, size and color) and orientations.

v Irregular text regions (non-rectangular)

Figure 6.1 shows few example document pages with complex layout which have been considered for the current work. In this section, we describe a new approach to decompose the document images with complex layouts, and label the segmented regions as text, graphics or background. Our approach defines a hierarchical architecture for segmentation. Colour information represents primary cue to separate the graphics and text regions from

(a) (b) (c)

Figure 6.1: Sample document images with complex layout

the background. Hence, we identify the dominant color planes in the image and process pixels corresponding to these planes using individual classifiers. Each classifier is trained by exploiting the local neighbourhood properties of the respective color plane. Finally, the contextual relationship across different color planes is learned by CRF to smoothen the plane-wise label assignment for final segmentation.

6.3.1 Scheme for Document Image Segmentation

The complete framework (Figure 6.2) proposed for text, graphics and background separation in documents with complex layouts consists of the following steps:

- Pre-processing
  - Colour analysis
  - Clustering
  - Feature extraction

- Initial segmentation
- CRF based smoothing

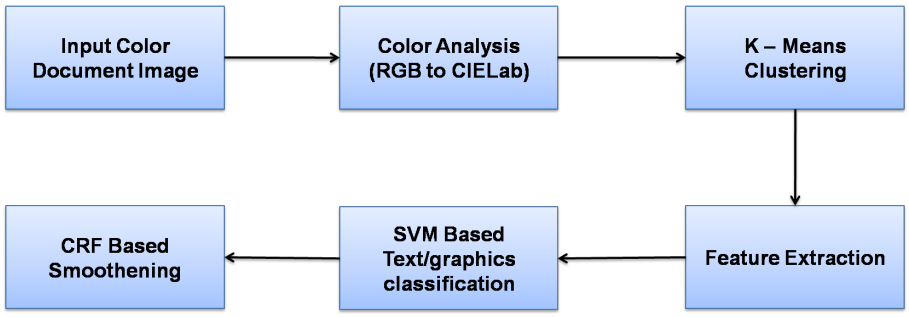


Figure 6.2: Architecture of the document segmentation framework

Colour Analysis

In many document images, different image components (text, graphics/image and background) appear in different colors. Additionally, in many cases the text overlaps pictures/graphics. Hence it becomes essential to extract local color information in uniform color space. Here we first convert the image from RGB color-space to a uniform color space such as CIELab space [279]. *Lab color space* has separate lightness and chroma channels that are approximately perceptually uniform and serve as a device independent color model. CIELab formulae is derived from CIEXYZ, therefore, conversion from RGB to CIELab will involve:

i Conversion from RGB to XYZ

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.412453 & 0.357580 & 0.180423 \\ 0.212671 & 0.715160 & 0.072169 \\ 0.019334 & 0.119193 & 0.950227 \end{bmatrix} \star \begin{bmatrix} R \\ G \\ B \end{bmatrix}$$

ii Converting XYZ to LAB Space

$$
\begin{aligned}
L^\star &= 116 f(Y/Y_n) - 16 \\
a^\star &= 500\,[f(X/X_n) - f(Y/Y_n)] \\
b^\star &= 200\,[f(Y/Y_n) - f(Z/Z_n)]
\end{aligned}
$$

Here X_n, Y_n and Z_n are the CIE XYZ tristimulus values of the reference white point (the subscript n suggests "normalized"), and f is defined as

$$
f(t) = \begin{cases} t^{1/3} & \text{if } t > (\frac{6}{29})^3 \\ \frac{1}{3}\left(\frac{29}{6}\right)^2 t + \frac{4}{29} & \text{Otherwise} \end{cases}
$$

Clustering

The initial segmentation is carried out by identifying the color modes using K-means clustering. Each cluster identifies a color plane representing pixels with similar color properties. Ideally, number of clusters should be equal to number of categories. However, the objective of the present work is to analyse document images having multi-coloured and overlapping text and non-text (graphics/picture) regions. Based on initial observation, we selected K = 8 as the optimum value, as very large K gives disconnected noisy regions decreasing the over-all segmentation performance. Figure 6.3 shows the different cluster planes obtained after K-mean clustering.

Feature Extraction

Features extraction for each cluster is discussed in the following section. Using the K-clusters, the image is decomposed in K-color planes. Subsequently, following local fea-

Figure 6.3: Colour plane identification in the image

tures are extracted from each color plane:

a. ***Gabor Features***: Texture features are based on the local power spectrum that are computed using 2D Gabor filter. Such filters are local and linear, characterized by the localization properties in both spatial domain and spatial frequency domain [122]. The 2D gabor filter is defined as follows:

$$g_{\lambda,\theta,\varphi,\delta,\gamma}(x,y) = Kexp(-\frac{x'^2+\gamma^2y'^2}{2\delta^2})\cos(2\pi\frac{x'}{\lambda}+\varphi) \tag{6.3.1}$$

Here $x' = xcos\theta + ysin\theta$, $y' = -xsin\theta + ycos\theta$ and λ is the wavelength of the cosine factor of the Gabor filter kernel, θ is the orientation, φ is the phase offset with φ = [0 $\pi/2$], δ is the standard deviation of Gaussian function and γ is the aspect ratio ($\gamma = 0.5$). A set of values for $\theta = [0, \pi/4, \pi/2, 3\pi/4]$ and $\lambda = [2, 4, 8]$ are used in our experiments that lead to generation 12 Gabor filters.

b. ***Edge Features***: Graphics/images are predominantly defined by high spatial frequency components whereas background regions are defined by low frequency components. In general, the textual content in document images are defined by frequency range lying between the high and low frequencies defining background and graphics regions. Image $I(x, y)$ is convolved using horizontal and vertical masks defined by Sobel, to obtain gradient components G_x and G_y. The gradient magnitude ($Grad_{mag}$) and orientation (θ) are computed at each pixel as:

$$Grad_{mag} = |\sqrt{(}G_x^2 + G_y^2)| \quad \text{and} \quad \theta = \tan^{-1}(\frac{G_y}{G_x})$$

Two *Local Gradient Histogram* features have been extracted:

b.i *Direction Histogram*: Edge Direction Histogram [273] feature extracts the statistical distribution of curvature found in text (in different scripts/languages) as compared to graphics/images. Most of the scripts have either horizontal and vertical straight lines or curly construction with almost no straight lines. On the other hand, for graphics regions edges distribution is random. The normalized edge direction histogram is computed locally for each pixel in 5×5 neighbourhood. The edges in the pixel neighbourhood are detected by convolving with horizontal and vertical Sobel masks. Subsequently, the edge directions at each pixel are computed using vertical and horizontal edges. Finally, we compute edge histogram using 12 quantization levels for our experiments.

b.ii *Magnitude Histogram*: Gradient computed over an image gives the information of directional change in intensity values of the image. Local histogram of the

gradient magnitudes is computed for every pixel in 5×5 neighbourhood. The gradient magnitudes are quantized into T (T = 10 in our experiments) bins to form a T-dimensional features vector corresponding to each pixel. The histogram normalized by its sum is applied for further classification purposes.

Initial Segmentation

Initial document segmentation is performed over each color plane by SVM based supervised learning. The classification performs a coarser segmentation of text, graphic/picture and background regions using the combination of features discussed above. The combination is done by concatenating various features. The multi-class SVM is architectured in Decision directed acyclic graph (DDAG) (Refer section 3.4.2). Figure 6.4 shows the resultant images for each cluster plane after SVM labelling. Green, red and blue pixels indicate picture/graphics, text and background regions respectively.

CRF based Post-processing

The labels from each cluster plane are propagated to a single plane as shown in figure 6.4. This plane is then subjected to CRF based smoothing, as the SVM based deterministic labels do not consider the contextual relationship across the pixel neighbourhood (Refer Appendix C for the detail of CRF). The results presented here are obtained with 5×5 neighbourhood. The results after CRF smoothing can be seen in figure 6.4.

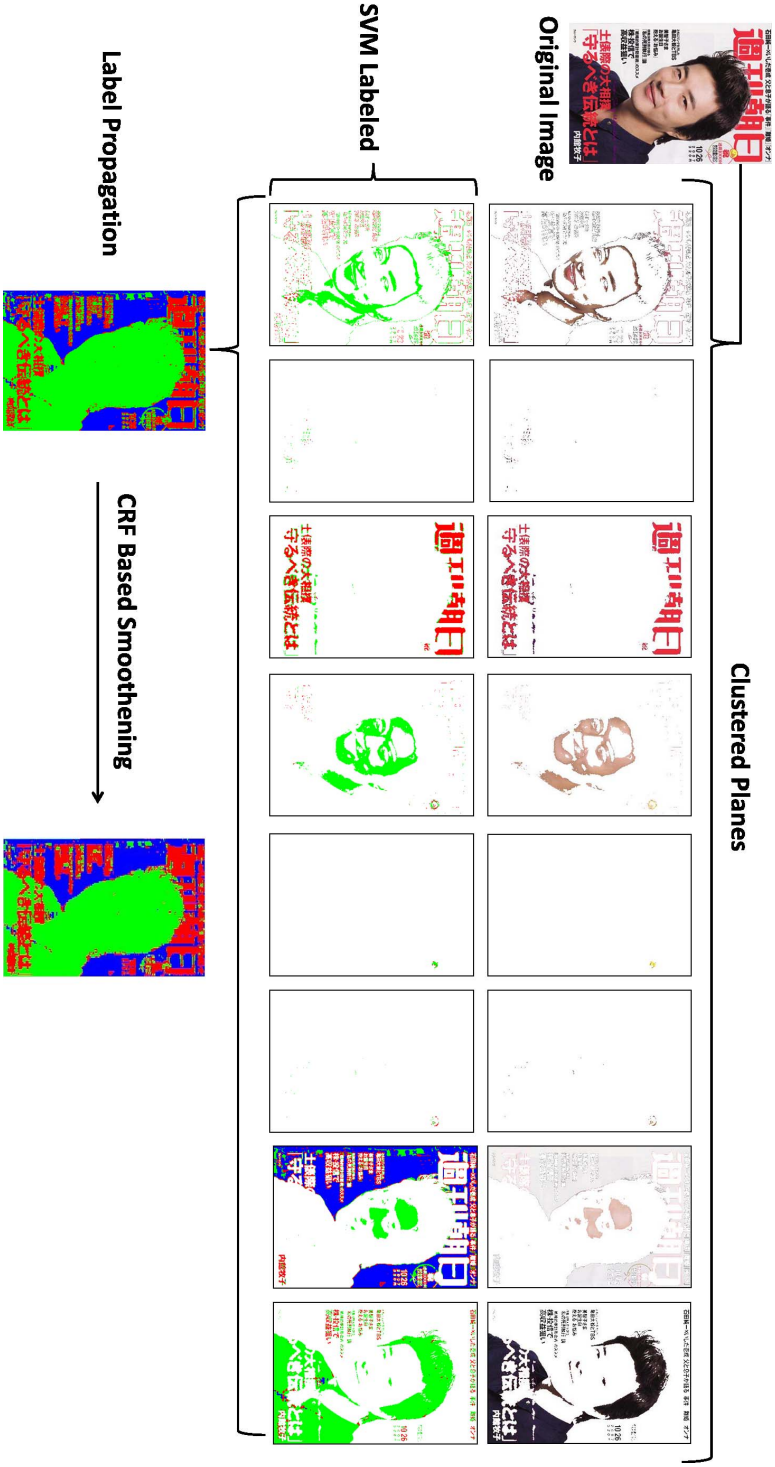


Figure 6.4: Text/graphics separation framework

6.3.2 Experimental Evaluation

The experimental dataset composed of document images consisting of the three classes: Text, Graphics/Image and Background. The collection comprises of document from magazines, articles and brochures, typically in non-manhattan layout. The collection primarily contains magazine front-pages which have text rendered in different styles, color and fonts and orientation and also overlaid on graphics. The approach is tested on 20 images with

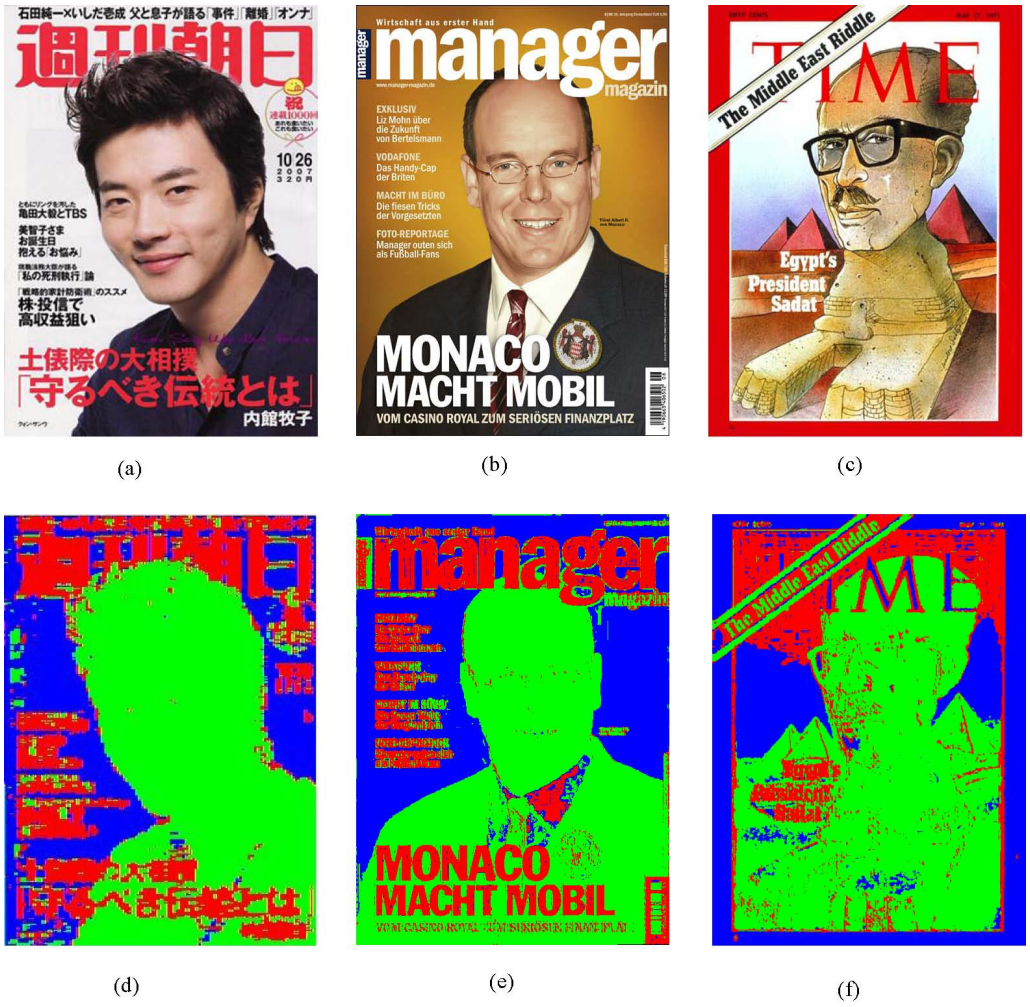


Figure 6.5: Original images and corresponding final segmentation

size varying from 400×600 to 900×1200. For all the images, the ground truth is prepared by manual segmentation. The training is performed by randomly selecting 4 images. The remaining images were used for testing. The experimental results have been presented as average of 3 iterations. The SVM classifiers for each color plane were trained by randomly sampling 5% of the total number of the training pixels corresponding to a color plane. We achieved 83.7% classification accuracy by SVM based classification. In general, for all the training images CRF based smoothing increased average classification accuracy by 4%-6%. Table 6.1 shows that after application of CRF, the final classification accuracy computed over three iterations, improved from 83.7% to 88%. Here the accuracy is measured as the percentage of label overlap for the document image. Figure 6.12 shows bleeding of text regions on the non-textual regions. In such cases, post-processing operation is required for accurate recognition.

Table 6.1: Final segmentation accuracies with SVM and CRF smoothening

Original Image	Final Segmentation Result	SVM	After CRF Smoothening
Figure 6.12(a)	Figure 6.12(d)	84%	86%
Figure 6.12(b)	Figure 6.12(e)	87%	89%
Figure 6.12(c)	Figure 6.12(f)	81%	83%

6.3.3 Multi-modal Retrieval of Document Images having Embedded Graphics

The text/graphics framework discussed in section 6.3.1 can be used for defining a multi-modal retrieval framework for document images having embedded graphics information. We propose multiple kernel learning based indexing discussed in section 5.4 for providing multi-modal retrieval having document images with graphics. The overall framework is presented in figure 6.6.

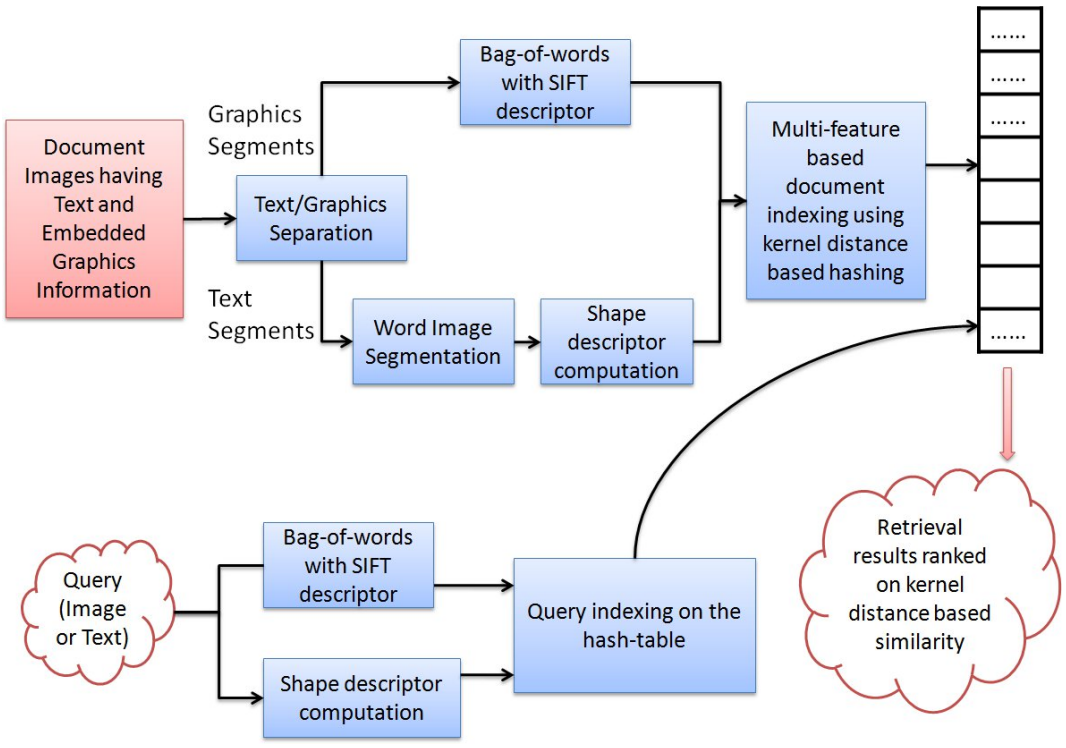


Figure 6.6: Multi-modal retrieval for document images having text and graphics

The framework generates unique document index by kernel distance based hashing, where optimal kernel for indexing is learned by combining kernels computed from

text, and graphics features. The framework computes query (Text or graphics) index on the hash table, and retrieves relevant documents by similarity search. The initial evaluation of the framework is performed on English magazine document collection having 192 scanned pages. After segmentation, the graphic components which are smaller than $1/50^{th}$ of the page size are filtered out. The segmentation process may generate noisy graphics regions. Therefore, bag-of-words computed with SIFT feature is applied for robust graphics representation (Details of SIFT and bag-of-words is presented in chapter 7). A visual vocabulary of 50 words is selected, therefore, resulting 50-d bag-of-words representation for graphics segments. The text segments are further segmented in word images as discussed in section 4.5. The segmentation resulted in 367 graphics image and 18871 word images. The set of linear and Gaussian kernels is used for forming the base kernel set corresponding to both type of information. For the evaluation, graphic segments having subjects such as faces, natural scenes, and buildings are selected. The validation query set X_v consisted 117 queries having 103 word object and 14 graphics objects. Final evaluation of the indexing is performed with query set X_q having 91 word and 8 graphics objects (Sample examples of graphics segments and word images are shown in figure 6.7). The base kernel set included linear and set of Gaussian kernels having variances from 2^{-5} to 2^5 with incremental step of 0.5 at exponent. The word images are represented by shape descriptor computed with following parameters: $\{m = 38, n = 36$, 1×6 partition$\}$. The genetic algorithm optimization is simulated for 100 iterations. The similarity search is done based on the kernel distance K_{dis} computes as $\sum_{i=1}^{n} w_i K_i$. The retrieval for hashing parameters $\{L = 50, k = 12\}$ achieved MAP score of 23.45% with graphics based queries,

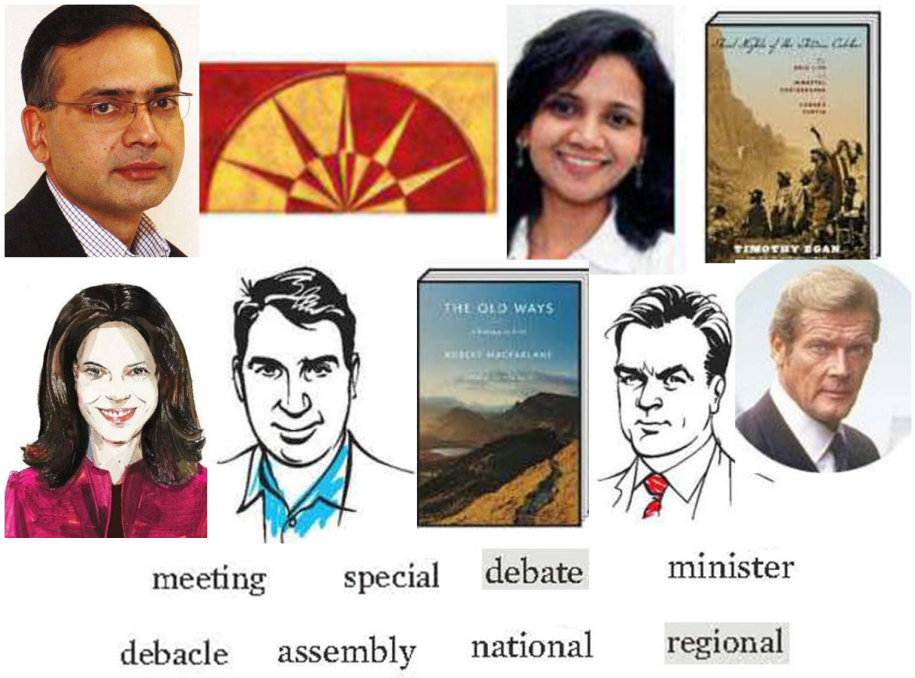


Figure 6.7: Examples from X_q

73.47% with text queries. The overall MAP score of 72.18% is achieved for query set X_q. The results establish that the segmentation approach can be successfully applied for developing a multi-modal content based retrieval framework for document images.

6.4 Script based Segmentation of Document Image

Script identification is an important step for automated text processing in document/natural scene images. Multilingual text is a common scenario in applications e.g. character recognizers, automated translation and document indexing and retrieval systems. We target general scenario of multilingual documents which have script variation at page, paragraph and word level. Such examples are common in dictionary, regional magazines, elemen-

tary level text books, official documents and examination questionnaires etc. Figure 6.1 shows an example of such document images having script variation at different levels. The document images in the figure have non-uniform distribution of English/Bengali and English/Devanagari words. The proposed framework helps recognition of such documents with multi-lingual text, as existing OCR engine can be directly applied without modification.

Large amount of work exists in automated script identification. Nevertheless, most of the work concentrates on either at page, line or word level. The identification at page level assumes uni-script document pages; therefore, limiting the scope and accuracy of the application. Additionally, the line and word level script identification is computationally expensive for practical applications. In this work, we consider the general scenario of multi-script text at page, paragraph and sentence level. We propose a novel script identification framework for multilingual document images by hierarchical combination of block and word level script identification.

Scripts can be discriminated based on their visual appearances. Single script text in a paragraph exhibit unique texture property which varies with scripts. Therefore, the proposed framework exploits the texture characteristics at paragraphs/blocks for script identification. Each paragraph is classified either an exclusive single script text or multi-script text block determined using the confidence score of the classifier. In this work, we present novel application of Relevance vector machine (RVM) for script identification at block level. The text blocks/paragraphs are extracted using the traditional recursive xy-cut algorithm followed by a split-merge approach. The split-merge technique combines sin-

gle text lines into blocks based on features such as interline spacing, alignment etc. The multi-script blocks are subsequently segmented into word images. We propose a novel application of rejection based classifier for script identification at word level. The classifier exploits the structural features of word images to learn the discriminative characteristics of different scripts. The framework therefore presents a computationally faster approach for script identification in document images than the direct identification at word level. We have considered bi-lingual document images with following combinations, *English/Hindi* and *English/Bengali* for testing the framework. The experimental dataset has been compiled from multilingual books, magazines and newspapers. The collection consists of real world examples having script variations at page/block/line level in non-uniform fashion.

6.4.1 Overall Framework

We propose a hierarchical framework for script identification in bilingual document images. The hierarchy consists of two prediction stages utilizing texture and structural features. The processing of subsequent stage depends on the classifier confidence for script prediction. The first stage exploits texture properties of the document image. We segment the text blocks from the image and identify the dominant script of the text. Blocks primarily represent the text paragraphs. The analysis of classification confidence decides the subsequent processing of blocks; where word level script identification is performed for blocks having multi-script text. At this stage, we apply script dependent rejection based classifier which is defined by cascaded combination of set of weak classifiers. The weak classifiers are learned for different false positive rates, i.e., it targets dominant script words as pos-

itives. The labelling process filters dominant script words at subsequent stages as *reject*. Consequently, the approach presents a fast classification method as majority of the words are rejected in initial stages of the cascade. The overall architecture of the framework is shown in figure 6.8. The details are discussed in the subsequent sections.

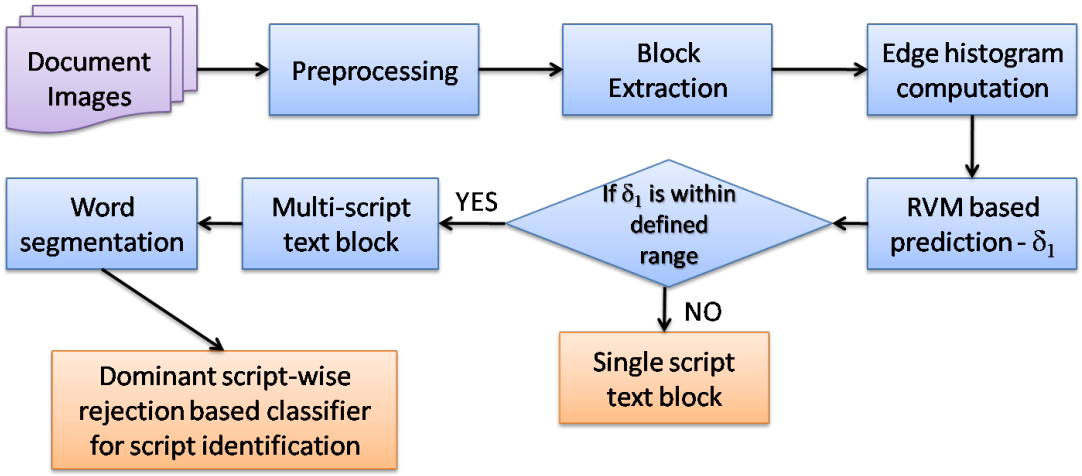


Figure 6.8: Architecture of script identification framework

Pre-processing

The document collection consists of gray-scale images scanned at 300dpi using a flatbed scanner. The conversion from gray-scale to two-tone image is performed by applying adaptive binarization routine discussed in [217]. The scanning process induces skewed documents images because of the mechanical error. Therefore, skew detection and correction is required to ensure proper orientation of the document images. Document image may contain noisy pixels and irregularities at the page borders or character boundaries decreasing the overall system performance. Adaptive median filtering is used for noise

reduction. After the pre-processing steps, document images are subjected to block level segmentation. Here, we assume that the images contain only textual content with different properties.

The segmentation algorithm described here segments document image into text blocks. The segmentation approach analyses the horizontal (X) and vertical (Y) projection profiles of the document image. The widths of valleys of horizontal and vertical projection profiles, helps in efficient tuning of segmentation parameters to detect appropriate vertical or horizontal separators. The position with least profile height is marked as a valid separator. This approach when applied to horizontal profile leads to line level segmentation. Traditionally, in many of image based applications feature extraction is preferred over blocks instead of single text line stripes. The argument for the approach lies in the fact that texture properties are more discriminatory when analysed over significant size of image block. In this case, the relative distribution of edges over a block provides a strong argument for the recognition. Therefore, we further merge single text lines to obtain blocks/paragraphs. Merging single text lines into blocks/paragraphs is based on the typographic parameters such as: alignment, interline spacing and white-space surrounding the blocks. Figure 6.9 shows the block segmentation for sample image. Blocks having same alignment, interline spacing are merged. In order to merge the indented blocks correctly, we analyse the white space surrounding the block. Using the analysis, the current block is merged with either the subsequent or previous text line or block. The word segmentation from text lines is performed by computing vertical projection profile. The discontinuities in the text lines are dealt by performing morphological operation before vertical profile

Figure 6.9: Block segmentation from example image

computation. We have performed dilation operation using line structuring element of length l. The observation on our dataset showed that $l = 7$ is the optimum value. The local minima's in the vertical profile define the possible word separators. These separators are projected over the original text lines for word segmentation.

6.4.2 Features Extraction

The document images corresponding to different scripts exhibit unique texture properties over text regions. The property is controlled by the frequency of curvature and straight

segments. Indian scripts have characteristic curvature distributions which help in visual discrimination. Therefore, edge direction based features and shape based features are employed to exploit the distribution of curvature for script identification. These features are font invariant and possess strong invariant characteristics to document degradation and noise. Some Indian scripts e.g. Devanagari and Bengali have the presence of horizontal line at the top of the word formation with vertical and horizontal strokes. Whereas prominent South-Indian scripts e.g. Tamil, Telugu, Malayalam and Kannada exhibit complex curved segments with no horizontal line and rare use of straight segments. In contrast to Indian scripts, English has dominant straight lines with certain amount of curves making the overall curvature distribution different. These script characteristics have motivated us to use global edge direction features at paragraph level and shape descriptor features at word level. Both the features are described as follows.

Edge Direction Histogram (EDH) We use the edge direction based features described in Section 6.3.1. The edge based features extract the statistical distribution of curvature found in image textures. Indian scripts have either horizontal and vertical straight lines or curly construction with almost no straight lines. The EDH features are computed on a global image patch containing the text to extract the orientation distribution of the script dependent curvatures. Therefore, the EDH essentially represents the global quantization of orientations in image patch. Whereas, HOG [62] discussed in Chapter 3 represents the local quantization of orientation computed in uniformly spaced smaller regions in image patch defined as cells, subsequently post-processed by local contrast normalization in

overlapped region defined by neighbouring cells. The edge direction feature is extracted in histogram with b bins. (b = 50 at block level) to obtain a b-dimensional feature vector. For handwritten character/digit recognition, a similar gradient based histogram has been introduced in [95].

Shape based Word Representation The word level script identification is performed by classifying the word images based on shape based feature discusses in Section 4.3.

6.4.3 Script Identification at Block Level

The block level script identification is performed by the application of RVM based classifier. RVM is the Bayesian alternative of state-of-the-art support vector machine. The primary motive for the application of RVM for classification is the advantage of probabilistic output. The probability output is used as the classifier confidence for selecting the block to process at next level. The detail of RVM is described in Appendix B.

Estimation of the Block Level Classification Confidence Interval

The block level classification gives probabilistic output which is used for decision on the subsequent level processing. The robustness of the framework is enhanced by defining a confidence interval over the prediction score of training blocks. The confidence scores are computed by means of five-fold cross validation. Mean of these scores are used for estimating the confidence intervals for each class. Interval estimates are desirable because the mean varies from sample to sample. Instead of a single estimate for the mean, a

confidence interval generates a lower and upper limit for the mean. The interval estimate gives the indication of uncertainty in our estimate of true mean. The narrower the interval, the more precise is estimate.

Confidence interval is expressed in terms of a confidence coefficient. Although the choice of confidence coefficient is somewhat arbitrary, in practice 95% interval is the most commonly used. The confidence coefficient is simply the proportion of samples of a given size that may be expected to contain the true mean. That is, for a 95% confidence interval, if many samples are collected and the confidence interval computed, in the long run about 95% of these intervals would contain the true mean [68]. The confidence interval is defined as:

$$CL = \pm \frac{1.96 \times \sigma_{X_j}}{\sqrt{NT}}$$

Where σ_{X_j} is the standard deviation of the confidence scores of X_j tests and NT is the number of tests. For the purpose of initial evaluation of the hypothesis, we computed confident interval for two datasets. First dataset consisted blocks, i.e., paragraphs having Hindi and English script words where 350 blocks belonged each of English and Hindi and 100 mixed blocks (Details of the dataset is presented in Section 6.4.5). Confidence intervals for pure English and Hindi blocks were estimated as $0.0094 < \delta_1 < 0.0247$ and $0.9763 < \delta_2 < 0.9859$ respectively. It was observed that for blocks having both English and Hindi words referred as the mixed blocks the confidence scores did not lie within the above defined confidence interval. Based on the above estimate, 97% mixed blocks were correctly identified. Table 6.2 shows some example mixed block and corresponding

scores. Also the final decision about the block based on the confidence scores is also listed. In the list, blocks with serial number 13 are 16 are missed out. Block 16 has few occurrences of English words. Corresponding confidence score signifies that it has Hindi as the dominant script because $\delta_2 > 0.5$. The odd occurrence of English words would contribute less in the overall error. Whereas block 13 has even distribution of words from both the script and would contribute more in overall error.

Table 6.2: Example text blocks, classification confidence score and final decision

S. No.	Sample Block	Confidence Score and final decision
1.	**हिन्दी अनुवाद**–'साहिब' एक पात्र है–इस कहानी का शीर्षक लेखक स्वयं है जो भारतवर्ष आया और कइ साल यहां रहा। वह इस देश में कई वर्ष रहा। उसका जीवन उतार-चढ़ाव से परिपूर्ण था। वह 4 वर्षों तक यहां रहा और कई तरह के परीक्षण किये। वह शिकार के क्षेत्र में बहुत ही प्रसिद्ध है। वह एक बहुत ही अच्छा शिकारी था और उसने शिकार के क्षेत्र में बहुत ही नये-नये परीक्षण किये। वह देश के दूर-दूर क्षेत्रों में गया। उसने एक बहुत ही प्रसिद्ध पुस्तक लिखी और उसका नाम था–'Man Eaters of Kumon' वह एक ऐसे पात्र से मिला जो बहुत ही निर्धन था। उसने उसे पांच-सौ रुपये उधार दिये। उसके व्यापार को जीवित करने का प्रयत्न किया। लालाजी ऐसा करने में सफल रहा। वह उसका पैसा वापिस करने आया। 'Sahib' ने उससे ब्याज के पैसे नहीं लिए। उसने पांच-सौ रुपया उससे स्वीकार कर लिया। इससे साहिब की दयालुता का प्रदर्शन होता है। उसने जरूरतमन्दों की सहायता की और उनकी रक्षा की। उसका सम्बन्ध उस समूह से है जो दूसरों की सहायता करते हैं। और जो सहायता करने के महत्त्व को समझते हैं।	0.9974016: Mixed
2.	**Explanation :** In the given lines, the writer says that workers made things for the rich. The poor could not buy these things. These were made by skilled workers with their personal interest. They loved their things. **हिन्दी अनुवाद**–ये पंक्तियां 'Mass Production' नामक पाठ में से ली गई हैं। इस पाठ के लेखक जी.सी. थार्नली हैं। इस पाठ में प्रचुर उत्पादक के गुणों और अवगुणों पर प्रकाश डाला गया है।	0.6040555: Mixed
3.	**हिन्दी अनुवाद**–हैनरी फोर्ड ने प्रचुर उत्पादन को कार बनाने की प्रणाली में इस्तेमाल किया। उसको इस दिशा में सफलता मिली। **Q. 7. Write a note on how the age of mass production was born.** **Ans.** Mass production was born with the introduction of big machines. These big machines started working at a big speed. **हिन्दी अनुवाद**–प्रचुर उत्पादन का प्रारम्भ बड़ी मशीनों के साथ हुआ। इन मशीनों ने रफ्तार से काम करना शुरू का दिया।	0.9988543: Mixed
4.	**Q. 9. How does mass production lower the price of an article ?** **Ans.** Mass production has reduced the cost of labour. New methods are developed to reduce the price of things. Moreover, the aim of every man of factory is to reduce the price. **हिन्दी अनुवाद**–प्रचुर उत्पादन में मजदूरी बहुत कम कर दी है। चीज़ों को सस्ता करने के नये-नये ढंग अपनाए जाते हैं। प्रत्येक माल बनाने वाला चीजों के दामों में कटौती करना चाहता है।	0.9750915: Mixed

Continued on next page

Table 6.2 – *Continued from previous page*

5.	वह अपने काम करने वालों से कहता है कि वह मौके के लिए सही कपड़े डाले। वह इस सालगिरह पर महिलाओं की उपस्थिति चाहता है। उसको भ्रम है कि महिलाएं किसी भी अवसर की शान होती हैं। पर किरिन महिलाओं को इस मौके पर नहीं आने देना चाहता। उसका मत है कि महिलाएं केवल अराजकता ही पैदा करती हैं। शिप्यूथिन सूचना देता है कि उसकी पत्नी किसी भी क्षण वहां आ सकती है। वह उसका वहां आना पसन्द नहीं करता। वह चाहता है कि वह अपने माँ-बाप के घर रहे। किन्तु टाटिना Tatiana वहां आ जाती है। वह उसे वहां से जाने के लिए कहता है। उसने इस अवसर के अनुसार कपड़े भी नहीं डाले। पर Tatiana वहीं पर आ जाती है। वह उसे वापिस जाने के लिए कहता है। पर वह नहीं जाती। Tatiana अपनी यात्रा के अनुभव सुनाना चाहती है। पर Tatiana बैंक से वापिस नहीं जाना चाहती। वह अपनी यात्रा के वृतांत अपने पति को सुनाना चाहती है। वह उसे घर भेजना चाहता है क्योंकि बैंक के भागीदार आने वाले हैं। तभी वहाँ पर अचानक Mrs.	0.9998869: Mixed
6.	**Q. 8. What is a worker's position in a modern factory, which has automatic machines ?** **Ans.** The worker has to operate the machine with skill. He must take care of the finished product also. **हिन्दी अनुवाद**–काम करने वालों को दक्षता से मशीन चलाना आना चाहिये और बाद में बने हुए माल को संभालकर रखना चाहिये।	0.3933831: Mixed
7.	**Vocabulary. Gloomy**–dark; **Shrieks**–screams, चीखें; **Vile**–bad; **Intention**–इरादे; **Reluctant**–not willing, अनिच्छुक. **Title : *Elizabeth Fry's Visit To Prison*** **Precis.** In January, 1817, Elizabeth Fry visited the women's ward at the Newgate prison. The wards warned her about the furious behaviour of the women prisoner. But Elizabeth was not detered and insisted on going inside. She refused to leave watch outside. The ward opened the gate unwillingly and Elizabeth entered the prison fearlessly.	0.3515619: Mixed
8.	**Vocabulary. Part and Parcel**–integral part, (अभिन्न अंग); **Consumer society**–that lives on consumption; **Huge**–big; **Roaring**–flourishing; **Stimulating**–exciting; **Hypnotising**–magical effect; **Impact**–effect.	0.0070918: Mixed
9.	It is because I know how sweet and happy and pure the home of honest poverty is, how free from perplexing care and from social envies and jealousies, how loving and united its members are in the common interest of supporting the family that I sympathise with the rich man's boy and congratulate the poor man's son. It is for these reasons that from the ranks of the poor so may strong, eminent self reliant men have always sprung and always must spring. If you will read the list of the immortals who were not born to die, you will find that most of them have been poor. **Vocabulary. Perplexing**–puzzling, confusing; **Jealousies**–ईर्ष्याएं; **Eminent**–famous.	0.0001823: Mixed
10.	**Vocabulary. Media**–medium, माध्यम; **Multiply**–to increase; **Oral**–u written; **Revolutionized**–completely changed; **Dimension**–a new area; **Her tage**–उत्तराधिकार; **Pollutes**–spoils; **Sensibilities**–tastes, attitudes; **Vital**–great **Title : *Press*** **Precis.** The printing press plays a role of great importance in this worl It spreads knowledge in time and space. Earlier knowledge passed on from o generation to the other. It was the method of ancient Gurus. Due to printi press knowledge has become one. But press is quite harmful when it prints lo type of material. But let's not blame press, the fault lies with man.	0.0000001: Mixed
11.	**हिन्दी अनुवाद**–सुकरात एक महान यूनान का दार्शनिक था। वह तर्कसंगत सोच-विचार में विश्वास रखता था। वह यूनान का जाना-पहचाना आदमी था। उसने गम्भीर और सोच-विचार का जीवन जिया। उसने यूनान के लोगों से तर्कसंगत व्यवहार करने के लिए कहा। सुकरात का ढंग प्रेरणा देने का था। उसका ढंग था–'Yes, Yes method'। इसका लक्ष्य था लोगों को अपने विचारों के अनुरूप बनाना तथा सोचने के लिए प्रेरित करना। इस ढंग का महत्व था कि यह तर्क पर आधारित था। सुकरात का ढंग तर्क का साथ था।	0.9866112: Mixed

Continued on next page

Table 6.2 – *Continued from previous page*

12.	**हिन्दी अनुवाद**–ये पंक्तियां 'Mass Production' नामक पाठ में से ली गई हैं। इनके लेखक हैं जी.सी. थार्नली। इस पाठ में प्रचुर उत्पादन के गुण तथा अवगुण प्रस्तुत किये गए हैं। इन पंक्तियों में लेखक कहता है कि प्रचुर उत्पादन बिना किसी दोष के नहीं है। इसकी अपनी ही सीमाएं हैं। वह कहता है कि कार्यरत व्यक्ति हमेशा खुश रहते हैं। काम से हमें सन्तुष्टि तथा प्रसन्नता प्राप्त होती है। कुछ भी न करना हमेशा ही दुखदायी होता है। अतः हमें अपने कार्य में व्यस्त रहना चाहिये।	0.98880266: Mixed
13.	**Ans.** A new method of manufacturing things was introduced. It was done to produce things on a large scale. 'Mass Production' was introduced in this connection. **हिन्दी अनुवाद**–वस्तुओं के उत्पादन के लिए एक नयी प्रणाली को अपनाया गया। यह प्रचुर उत्पादन के लिए किया गया। इसी सन्दर्भ में 'मास प्रोडक्शन' शब्द प्रचलित हो गया।	0.9848406: Hindi
14.	**हिन्दी अनुवाद**–ये पंक्तियां 'Garden City' नामक कविता में से ली गई हैं और इनके रचयिता हैं Sam O Nwaojigba. इस कविता में कवि कहता है कि Port Harcourt उस बाग की तरह जहां कोई फूल भी नहीं खिलता। इस स्थान पर निर्दयता है। यहां समय मृत को भी माफ नहीं करता। इस स्थान की दुखभरी कहानी है। यहां पर कुछ बिम्ब भुखमरी और मृत्यु की ओर इशारा करते हैं। दी गई पंक्तियों में कवि कहता है कि समय कमजोर व्यक्तियों पर ही मार करता है। यह मृत को भी माफ नहीं करता। इस कविता का परिदृश्य निराशा और निर्दयता से परिपूर्ण है। यह निर्धनता और अप्रसन्नता का देश है।	0.98704535: Mixed
15.	**हिन्दी अनुवाद–संदर्भ**–[जैसे कि part (*ii*)] दी गई पंक्तियों में कवि कहता है कि Garden City में वातावरण विषाक्त है। वह ऐसा महसूस होता है उन औरतों की तरह जिन्होंने छाती पर कवच पहन रखा हो। सांस लेना भी कठिन है। कवि उन औरतों का वर्णन करता है जो कि घोड़ों पर सवार हैं। यह नीरस मुस्कराहटें देती हैं। इन मुस्कराहटों में खुशी नहीं। मुस्कराहटें नीरस हैं और प्रसन्नता से वंचित हैं।	0.99434754: Mixed
16.	**हिन्दी अनुवाद**–ये पंक्तियां 'Management Speaks to the Graduate' नामक पाठ में से ली गई हैं। इस पाठ के रचयिता हैं Clarence B. Randall. इस पाठ में लेखक उसका मार्गदर्शन करता है वह नवयुवक जो औद्योगिक इकाइयों में मानवीय सम्बन्धों को समझना चाहता है। वह कहता है कि नवयुवक को मजदूरों की समस्याओं को समझना चाहिये। उनके साथ काम करना चाहिये उनके साथ सहानुभूति होनी चाहिये। उन्हें चाहिये कि अपने पर अनुशासन रखें और संतुलन भी। तभी मानवीय स्थितियों पर संयम रोपा जा सकता है। दी गई पंक्तियों में लेखक कहता है कि नवयुवक जो औद्योगिक इकाइयों से सम्बन्धित है उसे मजदूरों की समस्याओं को समझना चाहिये और मजदूरों के साथ काम करना चाहिये। उसे अपनी फैक्ट्री के उत्पादन को अच्छी तरह से जानना चाहिये। उन्हें पता होना चाहिये कि एक फैक्ट्री किस तरह काम करती है। इसी तरह नवयुवक औद्योगिक इकाई की समस्या को समझ सकता है।	0.9850017: Hindi
17.	**हिन्दी अनुवाद–संदर्भ**–ये पंक्तियां 'Bankers Are Just Like Anybody Else, Except Richer' नामक कविता में से ली गई हैं। इसके कवि हैं Ogden Nash। इस कविता में कवि कहता है कि बैंक बड़ी-बड़ी इमारतों में हैं। ये इमारतें अत्याधुनिक हैं। इसका कारण ग्राहकों को आकर्षित करना है। बैंक वाले पैसे से सम्बन्ध रखते हैं। वे पैसे का महत्व समझते हैं।	0.9908555: Mixed

Similar analysis is also performed on collection of blocks having English and Bengali words. The collection consisted of equal number of English and Bengali text blocks (Details of the dataset is presented in Section 6.4.5). For the estimation 400 blocks from each script is selected. RVM classifier is applied in five-fold cross validation approach.

The confidence interval for English and Bengali script words were estimated as $0.0017 < \delta_1 < 0.0318$ and $0.9763 < \delta_2 < 0.9911$. Based on the above estimate, the test on 61 mixed blocks showed that 96% of all the mixed blocks were correctly identified.

6.4.4 Script Identification at Word Level

The block level script identification may overlook the usage of multilingual words in text block/paragraph. RVM identifies the dominant script of the text block. Here, the probabilistic output represents the classification confidence. In this case, blocks having non-uniformly distributed mixed script words are also assigned correct labels. However, the classifier confidence identifies these blocks having exclusively single script text. Such blocks are processed at the next stage of the framework where word images from the text block are applied to a rejection based classification system. The classification confidence at the previous stage identifies the dominant script of the text. We train a pair of rejection based classifiers corresponding to both the scripts. The objective of pair of classifiers is to perform fast identification at word level.

Rejection based Classifier for Word Level Script Classification

The rejection based classifier applied for the purpose of script identification at word level is defined based on the cascaded classifier presented in [314]. The objective of cascaded combination of classifiers is to reject the dominating script words at the initial stage of the classification. The base classifiers at each stage are tuned with different threshold parameters to achieve 100% true positive rate at different false positive rates. The rejection based

approach therefore significantly reduces the identification time with robust classification performance. Here, four stage classifier hierarchy is used for prediction (Figure 6.10).

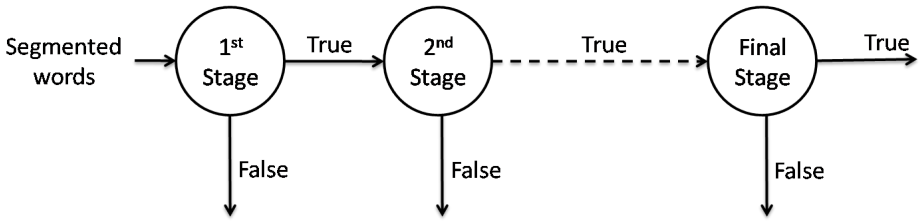


Figure 6.10: Cascaded classifier for word level script identification

The classification framework is similar to Adaboost classification. The shape based feature extraction process represent word images by high dimensional vectors. Adaboost adaptively selects the most discriminative and complimentary features from the training set, and formulates a strong classifier by learning based combination of several weak classifiers. The strong classifier applied for prediction task is defined as parametrized linear combination of several weak classifiers.

$$H(x) = sign < \{\sum_{m=1}^{T} w_m h_m(x) >\} \tag{6.4.1}$$

The parameter w_m defines the contribution of each weak hypothesis h_m for the prediction task. Weight parameters **w** are obtained by training as defined in the table 6.3.

6.4.5 Results and Discussion

The document collection for the validation of the proposed framework is compiled by scanning the supplementary books available for different type of courses e.g. Guide books and Language training books. The document images are scanned in gray-scale at 300

Table 6.3: Adaboost training algorithm

Consider the given the training data as $\{(x_n, y_n)\}_{n=1}^{N}$, where $x_n \in R^d$ and $y_n \in \{-1, +1\}$.
1 Initialize $\alpha_1^n = 1/N$. **For** each $m = 1, \ldots, T$
2 Train the weak classifier h_m using training dataset with example features weighted by α_m.
3 Classification error computation for h_m as $$\varepsilon_m = \frac{\sum_{i=1}^{N} \alpha_m^n \Delta\left(h_m(x_n), y_n\right)}{\sum_{n=1}^{N} \alpha_i}$$ $\Delta(a, b)$ represents the Kronecker delta function defined as 1 if a and b are equal.
4 Compute weight w_m for the weak classifier h_m as $$w_m = log\left(\frac{1-\varepsilon_m}{\varepsilon_m}\right)$$
5 Update the feature data distribution as $\forall n, \alpha_{m+1}^n = \alpha_{m+1}^n exp(w_m \Delta(h_m(x_n), y_n))$ **End**

dpi. A document image collection of 384 pages containing English/Hindi script and 526 pages containing English/Bengali scripts is used for experiments. The document images have been annotated as per the existence of the dominant script. For page level script identification, the document collections are partitioned as 70% for training and 30% for testing by random sampling. The feature details are discussed in section 6.4.2. The RVM based average classification accuracy over five iterations for both the document collections are listed in table 6.4.

Figure 6.11 shows the confidence scores obtained at block level script identification for few sampled blocks.

For English/Hindi document collection, we had 640 and 810 blocks belonging to

Table 6.4: Page level script identification

Document collection	Accuracy	Average confidence scores	
English/Hindi	99.12%	Hindi	0.9716
		English	0.0821
English/Bengali	99.38%	Bengali	0.9685
		English	0.0756

Hindi and English respectively. Similarly for English/Bengali document collection, we had 741 and 912 blocks for Bengali and English respectively. RVM training is performed by conventional 70/30 partitioning as discussed above. The classification results with average confidence scores are shown in table 6.5.

Table 6.5: Block level script identification

Document collection	Accuracy	Average confidence scores	
English/Hindi	99.06%	Hindi	0.9645
		English	0.0865
English/Bengali	98.93%	Bengali	0.9708
		English	0.0919

The word image collection details are as follows. Hindi, English and Bengali word image collection contains 4606, 8416 and 6718 examples respectively. The shape descriptor computation is performed with 40 distance and 36 angular bins by splitting the word image in four partitions i.e. $\{m = 40, n = 36, num\_parts = 4\}$. The results are presented in table 6.6. Few sampled blocks and corresponding script labelled images are shown in figure 6.12 to establish the validity of the framework. Finally, classification scores by combining all the stages is computed for overall performance evaluation.

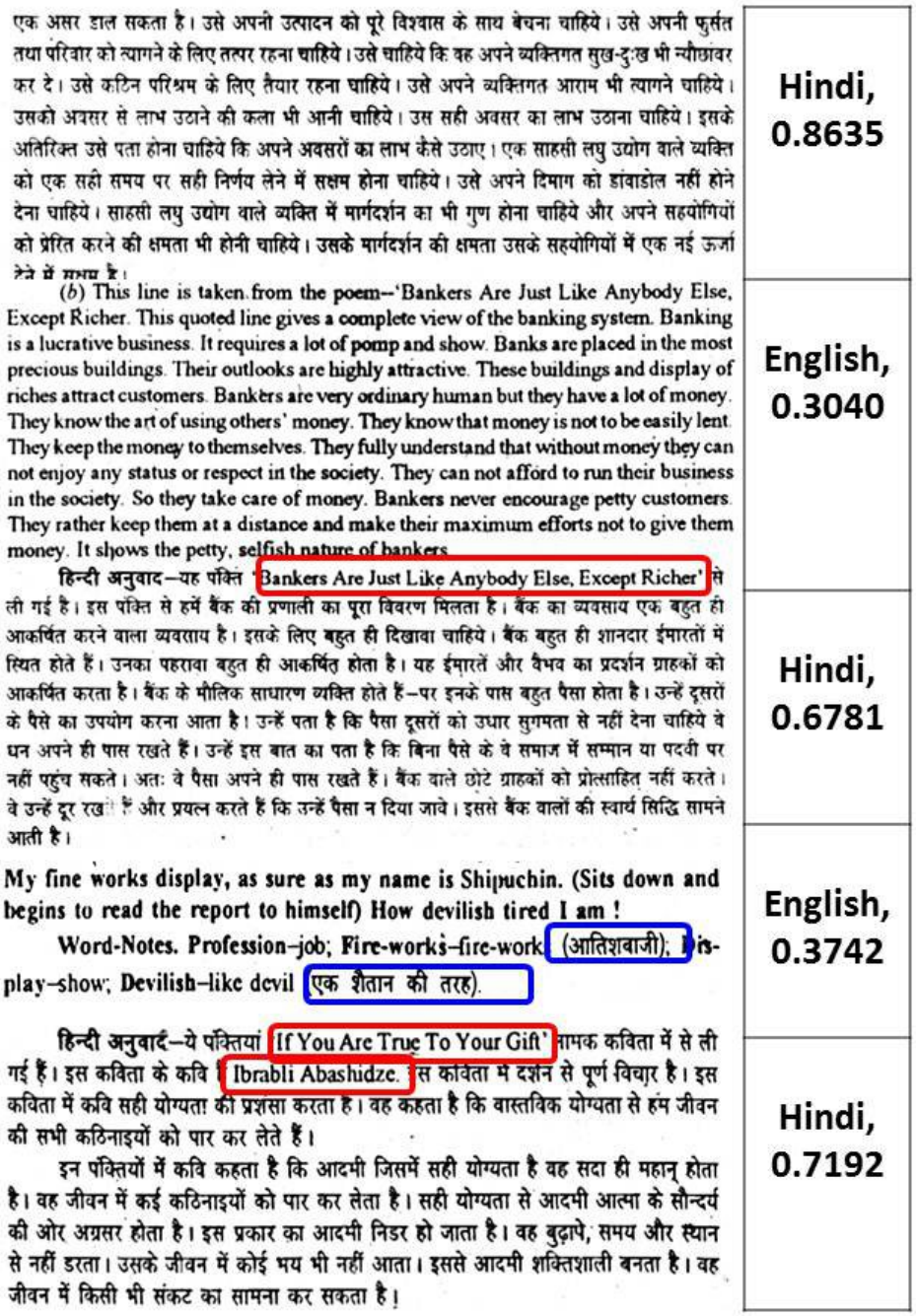


Figure 6.11: Block level script identification and corresponding confidence scores

First, the document image collections described in Section 6.4.5 are partitioned as 70/30 by random splitting. On the training set, confidence interval for page level classification

Table 6.6: Word level script identification

English/Hindi	English/Bengali
99.37%	99.11%

is evaluated by five-fold cross-validation. Using this confidence interval, the text blocks are generated. Five-fold cross-validation is applied for estimating the block level classification confidence interval. Which is subsequently applied for word level classification of multi-script images. Using this experimental methodology, overall word-level script identification accuracy of 98.17% is obtained for English/Hindi documents. Similarly, accuracy of 98.03% is achieved for English/Bengali documents.

6.5 LDA based Searching for OCR'ed Text

Conventional approach of similarity matching between the query keyword with database objects does not address the semantic requirement of querying. The topic based retrieval addresses the semantic requirement of querying by correlating the query theme with semantics of the documents. The topic modelling explores latent thematic structure of the document collection, and groups the documents having similar semantics [67, 135, 31]. The learning process extracts the latent semantic topics using the *tf-idf* (term frequency-inverse document frequency) based document representation. The document *tf-idf* computes the frequency of different 'terms' or 'words' with respect to the dictionary of unique 'terms' existing in the collection. Here, unique terms are the words which have atleast one occurrence in the document collection. In case of recognition errors, the term similarity

Figure 6.12: Script identification at word level

is inaccurately established, resulting in ineffective topic modelling.

In this section, we present topic modelling based document indexing framework for OCR'ed document images which proposes LDA modelling with term similarity computation. The objective is to define an indexing framework for OCR'ed , which can retrieve text documents with reasonable accuracy based upon keyword search, despite OCR's poor performance. The framework is defined without actually correcting the erroneous digitized text but exploiting performance characteristic of the OCR based on its character level confusion matrix. The topic model based indexing groups different terms occurring in the text document to discover hidden topics. The framework proposed here, exploits the knowledge of OCR's confusion matrix such that, the topic model captures the semantic relationship between terms modulo OCR's misclassification errors. For example, if an OCR consistently misclassify 'a' as 'o', then a term 'capitol' may be grouped with 'capital' for term frequency computation. The objective is to learn the topic model by alleviating the influence of recognition errors enabling semantic search with noisy OCR'ed text. This approach of semantic grouping performed by topic model based search will be more robust and realistic. Using the learned topic model, the topic based document representation is further applied for indexing using *Lucene* [306]. Here, we point that usage of string edit distance for recognition error correction has been well experimented. One such method available in Version 4 of *Lucene* [307] uses Levenshtein automaton which applies filtering methods for improving the finite state techniques for approximate search in large dictionaries.

6.5.1 Overall Framework: Document Indexing and Retrieval

The block level diagram of the proposed indexing framework is presented in figure 6.13. The procedure first learns a topic model over subset of ground truth data available for the OCR'ed document set. The learned topic model is used to index the OCR'ed document by inferencing. The *tf-idf* representation for recognized text is defined by modified edit distance based string similarity computation. Subsequently, the query is inferred for getting the distribution of different topics. Corresponding to the query topic distribution, we retrieve relevant documents by performing *Lucene* based search. The advantage of *Lucene* in combination with LDA is to provide fast search to retrieve documents corresponding to different topics. The detail of LDA is provided in Appendix D. Section 6.5.1 describes modified edit distance for inclusion of OCR's confusion characteristics for string similarity check. Subsequently, the details of indexing and retrieval framework is presented.

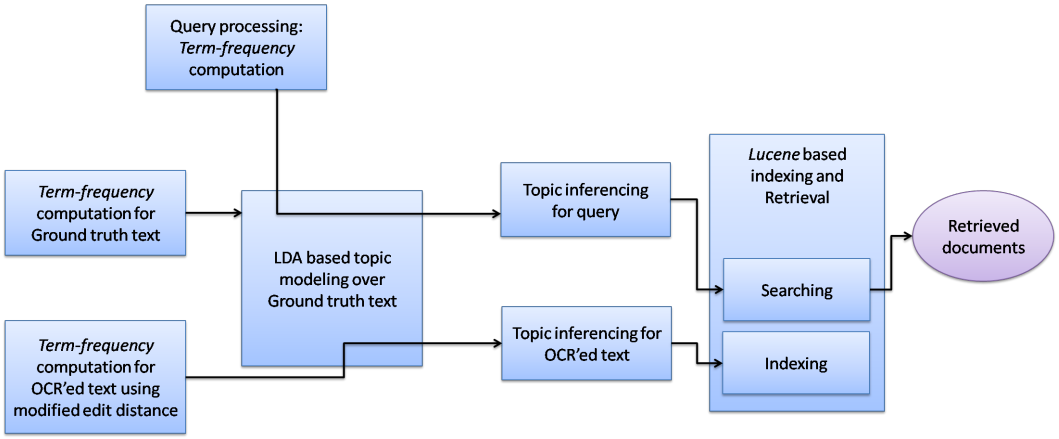


Figure 6.13: Document indexing and retrieval framework

Inclusion of OCR's Confidence Characteristics in Indexing

The confusion matrix of an OCR represents its confidence in recognizing various characters. The *tf-idf* representation of a document represents the frequency of different terms existing in the document. Here, we redefine the *tf-idf* computation by defining modified edit distance based string matching. The edit distance algorithm computes string similarity as the minimum number of insertion, deletion and substitution operations required for transforming one string to another (Refer section 4.7.2 for details). The algorithm assigns uniform penalty for all such operations. In the present case, we redefine the substitution cost based on OCR's confusion probability between i^{th} and j^{th} characters as:

$$\mathsf{Substitution\ Cost}(i,j) = 1 - \mathsf{OCR\_Conf\_mat}(i,j) \tag{6.5.1}$$

After modified edit distance computation between string Str1 and Str2, similarity is concluded by applying threshold over the distance. The summary of steps for string similarity computation using modified edit distance is as follows:

- Compute the modified edit distance between Str1 and Str2 using the substitution cost as in equation (6.5.1).
- Normalize the computed distance by length of longer string which is the upper bound of original edit distance.
- Conclude the string similarity by thresholding the normalized distance.

The *tf-idf* vectors for OCR'ed documents are generated using the above approach for string similarity. Subsequently, the topic distribution for documents is inferred by inferencing over the learned topic model.

6.5.2 Details of Indexing the OCR'ed Documents

The section lists the steps of indexing recognized document using LDA, and topic distribution based document retrieval using *Lucene*.

- Use the ground truth for randomly selected subset of text document collection for preparing the vocabulary of unique terms. The unique terms are subsequently used for computing *tf-idf* for ground truth documents.
- We learn LDA based topic model using the *tf-idf* of ground truth which explores the semantic grouping of different document terms.
- The *tf-idf* vectors for OCR'ed documents are computed with vocabulary of unique learns generated from the ground truth collection. The word string similarity is established by computing modified edit distance as discussed in section 6.5.1. The selection of threshold for concluding the similarity/dis-similarity of strings is done based on the statistical analysis of OCR output. We varied the threshold within the range of average recognition error $\pm$ 10% of the variance of recognition accuracy. The final threshold selection requires tuning so that a unique term in the vocabulary is not concluded similar to highly dissimilar words.
- The next step converts *tf-idf* vector for all OCR'ed documents in topic distribution by inferencing using the learned LDA model.
- The topic vectors for OCR'ed documents are further indexed using *Lucene* platform. Here, numeric field mechanism is used instead of traditional term based indexing. Each topic is considered a numeric field for a document, and topic-weight (probabil-

ity) is considered its value. If we have k-dimensional topic distribution, document d_i is converted into vector t_i as $\{t_{i1}, t_{i2}, ..., t_{ik}\}$. Each t_{ij} is added as numeric field to corresponding *Lucene* document.

Retrieval Framework

It is always desirable that documents having topic distribution same as the query should always be retrieved irrespective of recognition errors. The retrieval process begins with conversion of query into topic as $\{q = t_{i1}, t_{i2}, ..., t_{ik}\}$ by learned the LDA model using ground truth. Subsequently, topic vector for q is searched in *Lucene* indices using numeric range query. The numeric range query compares the attributes within a range for scoring the documents. Here, the selection of range requires parameter tuning. Using a small range may neglect many relevant results, therefore resulting low recall rate. The large range assigns same score to many documents, therefore, reduces the precision of retrieval. However, numeric scores do not always help in distinguishing, or ranking the documents falling in same range. Therefore, cosine similarity based score to rank the relevant documents returned by *Lucene* search. The process provides a fast retrieval platform as *Lucene* based search confines the set of documents ranked by cosine based similarity measurement.

6.5.3 Experimental Validation

In general, the OCR technology for Indian scripts is not standardized so far, therefore sample document images of Devanagari script are used for experimental evaluation. The

collection is prepared as part of consortium project funded by Government of India [15]. The experimental collection contains 600 document images. The documents are digitized by an OCR which achieved 77.8% accuracy at character level. The retrieval framework is evaluated for 61 synthetic queries having 2 to 4 words. The edit-distance threshold is selected as 0.1 for generating the *tf-idf* for recognized documents. For performance measurement of the framework, we learn the topic distribution for corresponding ground truths using learned LDA model, and index over *Lucene*. The evaluation is performed by measuring the overlap of retrieved documents for a given query, i.e., overlap of the OCR'ed text document and corresponding ground truth document. The results for different number of topics and variation in range queries are presented in figure 6.14. In case of range parameter value as 0.5 for OCR'ed text, MAP measures of 49.45% and 53.81% are achieved for 20 and 40 topics respectively. In case of corresponding ground truth with similar range parameter selection, MAP measures of 54.56% and 59.66% are achieved for 20 and 40 topics respectively. The results establish the validity of proposed framework. The overlap between the retrieved results improves significantly with increase in search radius. It is always desired to have all the relevant documents at top of ranked results. In this context, the framework achieves encouraging results as the overlap is maximum in case of top 3 relevant results. However, the application of *Lucene* for indexing the document topic vectors restricts the semantic nature of topic modelling based retrieval. In the present context, LDA based topic modelling is applied for generating latent topic based document distribution with invariance to recognition errors. However, the retrieval phase does not follow the conditional probability estimate for defining the relevance with

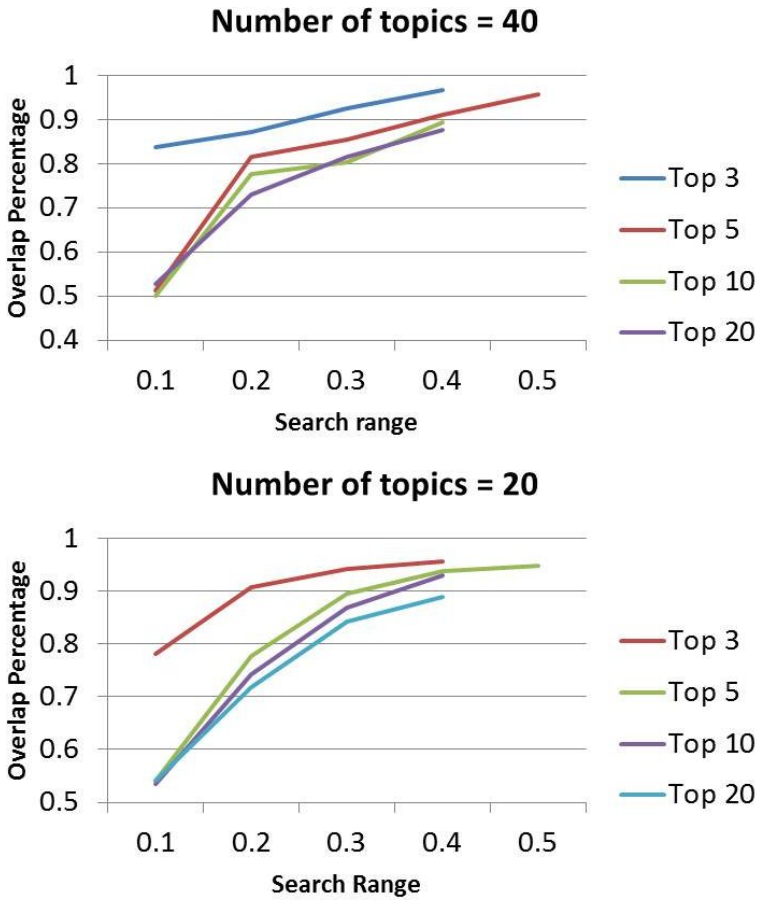


Figure 6.14: Percentage overlap for given query set over retrieved documents from ground truth and OCR'ed text

the query.

6.6 Word based Multi-modal Document Image Indexing

Section 6.5 discussed method for retrieving the noisy text documents having recognition errors. However, the retrieval performance is not guaranteed in case of significant errors. We propose improvement in retrieval performance of such documents by defining a multi-

modal document indexing scheme. The framework follows multiple kernel learning based indexing discussed in section 5.4 which is applied to generate a unified indexing space of noisy text documents with corresponding document images. Figure 5.4 shows the block diagram of proposed indexing scheme. The following feature representations are used for defining the word based document indexing scheme.

- The image form of documents i.e. document images are segmented in collection of word images. The word images are represented by shape descriptor (Section 4.3) for generating the document indices.
- In section 6.5, topic model based indexing and retrieval framework is proposed for text documents. The framework presented novel method for noise invariant topic assignment to document terms which uses the recognizer's confusion characteristics in topic learning. This topic distribution for text terms i.e. $p(\mathbf{z}|\mathbf{w})$ is used to represent the recognized documents in the unified indexing scheme using kernel distance based hashing.

The topic distribution based feature set represents distinct statistical characteristics. Therefore, the kernel function based similarity should compare probability distributions. In this work, symmetric Kullback-Leibler (KL) divergence based kernel presented in [219] is applied for kernel matrix computation using topic distribution for text terms. The kernel function is defined as:

$$K(\mathbf{w}_i, \mathbf{w}_j) = K(p(\mathbf{z}|\mathbf{w}_i), p(\mathbf{z}|\mathbf{w}_j)) = \exp^{-\alpha D(p(\mathbf{z}|\mathbf{w}_i), p(\mathbf{z}|\mathbf{w}_j))+\beta} \tag{6.6.1}$$

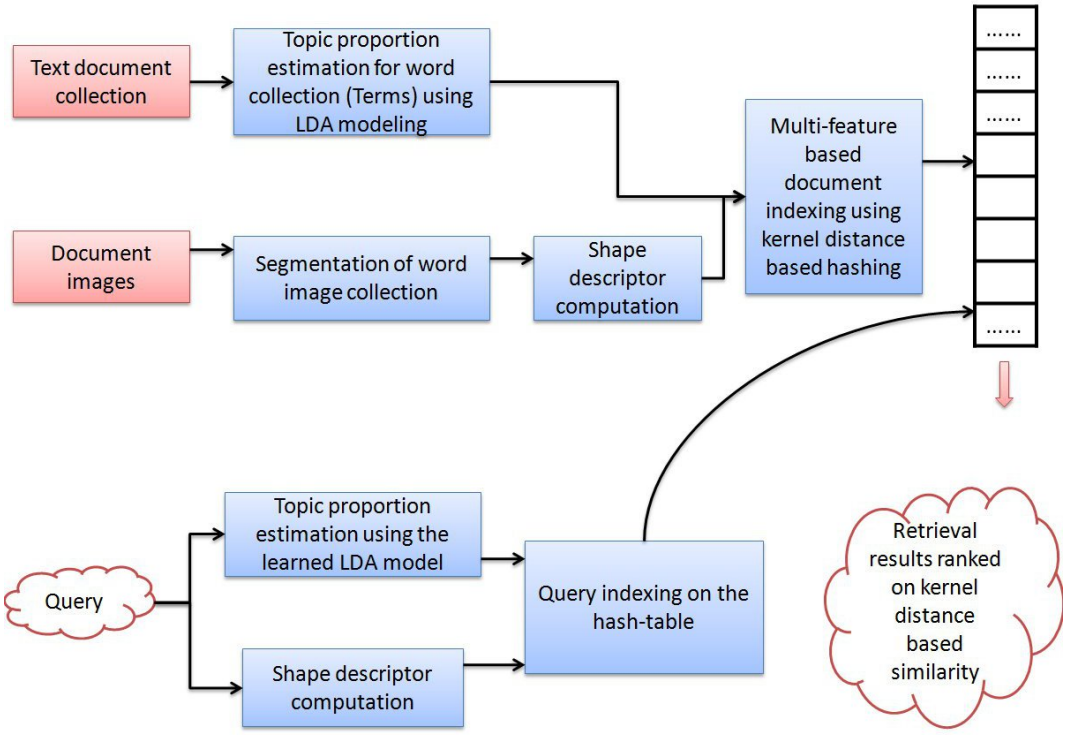


Figure 6.15: Word based multi-modal document image indexing

The α and β are scale and shift parameters which are supplied in advance. A range of these parameters is selected to define corresponding base kernel set. Distance D is the symmetric KL divergence between $p(\mathbf{z}|\mathbf{w}_i)$ and $p(\mathbf{z}|\mathbf{w}_j)$ as follows:

$$D(p(\mathbf{z}|\mathbf{w}_i), p(\mathbf{z}|\mathbf{w}_j) = \int_{-\infty}^{\infty} p(\mathbf{z}|\mathbf{w}_i) log(\frac{p(\mathbf{z}|\mathbf{w}_i)}{p(\mathbf{z}|\mathbf{w}_j)}) d\mathbf{z} + \int_{-\infty}^{\infty} p(\mathbf{z}|\mathbf{w}_j) log(\frac{p(\mathbf{z}|\mathbf{w}_j)}{p(\mathbf{z}|\mathbf{w}_i)}) d\mathbf{z} \quad (6.6.2)$$

Experimental evaluation of the framework is discussed in next section.

6.6.1 Experimental Evaluation

The proposed multi-modal document image indexing framework is evaluated on document collection of Devanagari script. The document collection described in section 6.5.3 is used for evaluation. After initial filtering of terms having less than 3 characters, the OCR'ed text documents consisted of 18895 terms. Following the word segmentation and filtering process discussed in section 4.5, corresponding document images generated collection of 18174 word images. The validation query set X_v consisted 76 terms/words common to both type of documents. Final evaluation of the indexing framework is performed for query set X_q having 83 query words with length varying from 4 to 9 characters common from both collection. The base kernel set for topic distribution based representation included kernels having parameter $\alpha = \{0.1, 0.5, 1, 2, 10, 5, 10\}$. The shift parameter β is set as zero. For shape descriptor based representation, the base kernel set consisted of linear, and Gaussian kernels with variance = {1, 4, 10, 20, 100}. Shape descriptor for word images is computed with following parameters: $\{m = 50, n = 45$, 1×4 partition$\}$. The GA optimization is simulated for 100 iterations. The kernel space similarity is used for ranking the retrieval results by computing based kernel distance as $K = \sum_{i=1}^{n} w_i K_i$. First, indexing and retrieval performance is evaluated for independent collection. The results for different hashing parameters are shown in table 6.7. The retrieval of text documents depends on the effectiveness of topic distribution based term representation. The dimensionality of feature space i.e. number of topics ($|\mathbf{z}|$) is therefore required to be carefully selected. In the next step, both types of documents are combined following the framework described in

Table 6.7: Retrieval results on different types of documents

Results on OCR'ed documents								
	L = 40, k = 12		L = 40, k = 20		L = 50, k = 12		L = 50, k = 20	
	MAP	Avg Comp	MAP	Avg Comp	MAP	Avg Comp	MAP	Avg Comp
$\|\mathbf{z}\|$ = 25	47.76	2302	42.25	1845	50.21	2734	45.46	2102
$\|\mathbf{z}\|$ = 50	62.37	2184	55.11	1742	64.54	2605	57.69	1955
$\|\mathbf{z}\|$ = 100	61.43	2095	53.93	1717	63.89	2511	56.26	1886

Results on text document in image form							
L = 40, k = 12		L = 40, k = 20		L = 50, k = 12		L = 50, k = 20	
MAP	Avg Comp	MAP	Avg Comp	MAP	Avg Comp	MAP	Avg Comp
75.33	1576	71.49	989	76.86	1648	72.58	1071

figure 6.15. The retrieval experiment with individual type of document collection showed that $|\mathbf{z}|$ = 50 achieved best results. Therefore, the topic distribution learned for $|\mathbf{z}|$ = 50 is subsequently used text term representation. The base kernel set for multi-modal indexing is formulated by ORing both the based kernel sets. The evaluation results for different hashing parameters are presented in table 6.8. The results show $4.42 \sim 6.20\%$ increase in MAP measures for different hashing parameters. The results establish that proposed indexing framework can significantly improve the retrieval accuracy on noisy text documents by generating unified indices with corresponding document images.

Table 6.8: Retrieval results on combined collection (OCR'ed documents and text documents in image form)

L = 40, k = 12		L = 40, k = 20		L = 50, k = 12		L = 50, k = 20	
MAP	Avg Comp	MAP	Avg Comp	MAP	Avg Comp	MAP	Avg Comp
80.67	2854	75.91	2332	82.06	3205	77.14	2662

6.7 Conclusions

The chapter presented novel methods for integrating information for multi-modal document retrieval in multi-modal, multi-color and multi-script scenario. The chapter concentrates on the extraction of multiple modality regions and script information of textual segments. First, an approach for the segmentation of text and graphics modalities from document images is presented. The approach presents a framework for general scenario of overlapped text/graphics segments by combining the unsupervised and supervised learning. The method efficiently exploits the local contextual information in image and color space for segmenting different regions. The efficacy of method is demonstrated on class of multi-modal documents covering magazine cover pages depicting text and graphics segments in random fashion. Subsequently, we demonstrated the application of segmentation framework for multi-modal retrieval of document images.Information of the script of textual content is required in advance for automated digitization of document images. Novel framework for solving the script identification problem in bi-lingual documents is presented. The framework presents a robust and fast approach by unifying page, block and word level script identification which addresses the practical requirement in applications having documents with non-uniform distribution of languages/scripts. The successful application of structural feature with rejection based classifier is demonstrated for word level script identification. The extensive evaluation of the proposed concept is presented on two document collections. We have proposed topic learning on noisy text documents by modified edit distance. Application of the concept is shown for indexing and retrieval of text

collection prepared by inaccurate OCR technology. The work presented novel application of Lucene for topic based search on the text documents. The efficacy of the proposed framework is demonstrated over Devanagari script document collection. Subsequently, the topic learning using modified edit distance is applied for word based multi-modal document image indexing. We demonstrated that learning based feature combination of topic distribution of text terms, and shape descriptor of corresponding word images can significantly improve the overall document retrieval performance.

Chapter 7

Concept Learning for Multimedia Content Handling

7.1 Introduction

Content based information retrieval systems manage multimedia documents by extracting the low-level features without exploiting the conceptual model of information present in documents. The query requirement is expressed using manually created examples. However, the correlation between query image description by low-level features, and semantics of the query is always questionable: a problem referred as semantic gap. The text based retrieval expresses the query requirement by keywords representing high-level semantic concepts. These keywords relate to different multimedia documents based on the content and context level correlation. For sufficiently large concept vocabulary, most of the queries can be accurately expressed which is not possible in content based retrieval.

The concept based retrieval first requires the identification of different concepts underlying the document collection. In this direction, the collective effort by multimedia researchers and developers have culminated in LSCOM [224], which presented annotated collection of different videos by more than thousand high-level concepts related to people, objects, locations, and programs. Traditionally, high-level concepts from multimedia have been learned by supervised learning procedure which consider the known examples having similar semantic meaning corresponding to one conceptual class. The existing methods in this direction have primarily used single feature based learning for mapping the low-level features to high-level semantic concepts. Identification of domain-specific concepts is a challenging task as manual detection is immensely cumbersome and requires domain expertise for selection of meaningful concepts. Also, the generalization of methods across domains is difficult. In many cases, concept identification require exploration of latent semantics of data for analysis. Documents having multi-modal components pose another level of challenge for conceptual tagging. Considering, uniqueness of the feature space of different media forms, concept level fusion of multi-modal information appears more logical approach for application dealing with multi-modal content.

In this direction, the chapter presents novel methods for concept based multimedia content analysis and modelling. First, we describe concept learning frameworks for recognition and retrieval problems by the application of multiple features. Subsequently, we describe methods for multi-modal multimedia data analysis in which the latent semantic concepts belonging to the document collection have been learned based upon media features. In summary, the major contributions in the chapter are as follows.

- We propose a concept learning framework which uses multiple feature based low-level content representation. The framework uses multiple kernel learning (MKL) based classification discussed in section 3.4.2. We have proposed new image feature using the texture property. The texture feature is combined with SIFT descriptor for high-level concept based annotation of Indian classical dance posture images. We have also evaluated performance of MKL for recognizing the LSCOM concepts. Subsequently, we explore how the MKL framework presented in section 5.3 can be used for concept based image retrieval.
- We have proposed a scheme for establishing conceptual linkage across documents of multiple modalities as well as multi-modal components of the same document using unique combination of learned generative and discriminative probabilistic inferencing.
- We have used similar combination of generative and discriminative probabilistic reasoning for recognition of multi-modal concepts. As case study, we have considered event recognition in sports video and concept detection in multi-modal documents.

7.2 MKL based Concept Learning

In this section, we describe novel image concept learning framework by feature combination. The applicability of the framework is shown for annotating the class of images exhibiting postures for Indian classical dance domain. The concepts here represent the broad level categorization of dance postures which are subsequently applied for develop-

ing concept based domain specific ontology of dance videos. In general, large number of concept classes can be identified for semantically similar images. Semantically similar images, i.e., images representing same object or concept, may exhibit completely diverse visual appearance. Effective image representation for characterizing diverse visual image properties addresses semantically similar visually diverse images by relating the high level concepts to image content. Several efficient low-level feature extraction algorithms are available which can capture subtle variations in colors, color layouts and textures of images. However, the semantic gap between low level image features and high level concepts (annotation tags) is still a open problem, low-level features can not describe the complete image semantics. We propose combination of multiple features for improving the image annotation accuracy. The experiments have been performed on posture image collection corresponding to Indian classical dances (example images shown in figure 7.1). The im-

Figure 7.1: Chawk posture in Odissi dance style

ages contain dense multimedia information in terms of color, illumination and view point etc. Annotation of these images using posture based concepts is challenging as it requires the invariance to camera view-point, dancer's position, costume color, and background

across the examples. Figure 7.1 shows the *chawk* posture of Odissi dance sequence. Here, the conceptual model of dance postures has to be view point, position and color invariant. For such application, an image representation invariant to color, position, and viewpoint is required. We have presented texture feature for sufficient characterization of visual properties of posture images. Single feature based learning can not exploit the diverse visual image properties represented by high dimensional multi-modal features. We propose MKL based annotation which applies learning based combination of diverse feature representations for distinguishing different posture concepts. The classification architecture discussed in section 3.4.2 is applied for the task. The following section describes the features used in this work.

7.2.1 Feature Description

The following features have been used for concept based image annotation task of dance posture images.

SIFT Feature[192]

The SIFT feature represents an image by collection of local feature vectors. Each feature vector is associated with a key-point on the image. The initial step of key-point identification is done by analysing the image scale space at multiple scale. Key points are the local maxima/minima points, obtained by applying Difference of Gaussian (DoG) operator at image scale space. Selection of maxima/minima locations helps in achieving rotational invariance in key point selection. The identification of key points is performed by generat-

ing the image pyramid by re-sampling between each level. The detection of maxima and minima is performed by comparing each pixel in the pyramid with its 8-neighbours. First, the comparison is performed at same level of the image pyramid. If the pixel is maxima or minima at this level, closest pixel location at next lower level is identified by performing 1.5 times re-sampling. If the pixel remains lower or higher than the closest pixel location and its 8-neighbours, we compare the pixel with closest pixel location and its 8-neighbours at a level above. If the pixel corresponds to local maxima or minima location point, we consider this pixel as key point.

The feature vector computation corresponding to each key point is done by computing the gradient and orientation information for each Gaussian smoothed image for each level. Each key point is assigned a canonical orientation, which is computed as the peak of local image orientation histogram. Once the key points at different image scale space are identified and orientation values is assigned to all of them, key point feature vector describing local image region is computed. First, the local image region around the key point on the image closest to key point's scale is represented by multiple images representing each of a number of orientations. Lowe [192] has used $16\times$ 16 image region around the key point. 8 orientation planes are used, with each plane sampled over a sub region of $4\times$ 4 pixels. Here the orientation of pixels is computed with respect to orientation of key point. Thus, the descriptor corresponding to each key point at lowest scale is a vector of $4\times$ $4\times$ 8 elements.

Local Texture Feature

Texture feature is an important cue for image analysis which has been extensively applied for content based image retrieval. An image texture is defined as set of local neighbourhood properties of the gray levels of an image region. The proposed texture feature describes the image content at multiple resolutions having spatial as well as frequency information.

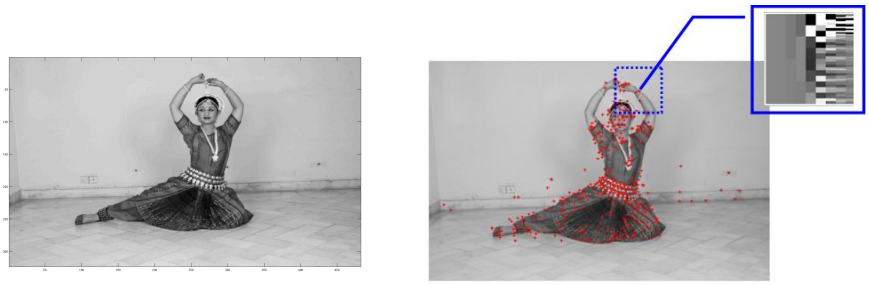


Figure 7.2: Distribution of key points on the image in left and local texture feature for a point

Fundamentally, an image could be considered as mosaic of different texture regions. In our texture based feature representation, we consider image as mosaic of overlapping texture regions. The regions are identified as the neighbourhood of salient key points. The key point location identification follows the SIFT key point identification as discussed in section 7.2.1. We have considered Haar wavelet response to describing the texture property in the local neighbourhood of key points. A 33×33 pixel neighbourhood around each key point is identified for wavelet response computation (Figure 7.2 shows key points, and wavelet response corresponding to a sample key point). The noisy key points, i.e., points

for which neighbourhood region crosses the image boundary have been neglected. Here, the approximation coefficients obtained after 2-scale decomposition of neighbourhood region have been considered, as the fine scale wavelets capturing high frequency details are inefficient in characterization of object details in dance images. However, the fine scale wavelets could be considered for different applications. For enhancing the effectiveness of wavelet response against intensity value transitions, we normalize coefficients value with mean of approximation coefficients in the local neighbourhood. Global texture features have been preferred for texture based image representation [34, 13, 352]. However, proposed method of local wavelet response computation using SIFT key-points incorporates scale, illumination, and view-point invariance. The bag-of-words model is further applied for quantizing the local wavelet responses, which defines robust image representation with invariance to spatial placement of object in the image. Additionally, it gives us constant size feature representation which could be conveniently applied to different learning algorithms.

Indicator Vector Computation

The SIFT and local texture feature characterize the image by set of local feature vectors. The variable number of local feature vectors for different images poses difficulty for MKL learning which require feature representation of constant dimension. Using both the feature extraction methods, many key point neighbourhoods of different images, and the same image are similar in terms of feature value. Therefore, the set of local feature vectors can be converted into constant dimension vector by feature space quantization. Following the

term-document representation used for text documents [135], each key point neighbourhood is associated to a visual word. The fixed dimension vector for images referred as bag-of-words representation is generated using a visual vocabulary. The simplest method to generate visual vocabulary is use of k-means clustering on the local feature vectors of training images. The clustering generates a visual vocabulary having k code words computed as center of learned clusters. The bag-of-words based representation generates a constant size image representation which is invariant to small variations in local feature vectors corresponding to similar image regions. The indicator vector is computed by mapping each visual word to one code word. Formally, the image is represented as vector

$$\overrightarrow{I} = \{sw_1 : sw_2 : \cdots : sw_b\}$$

Here b represents the vocabulary size and, sw_i represents the frequency of i^{th} *code word* in the image.

7.2.2 Annotation Model Architecture

The proposed annotation model is based on MKL based large-class labelling architecture presented in chapter 3. The architecture arranges the set of pair-wise MKL classifiers in DDAG formulation to perform robust and fast recognition. The class labels in this case are the posture concepts identified in the image collection. The classifier architecture is represented in the figure 7.3.

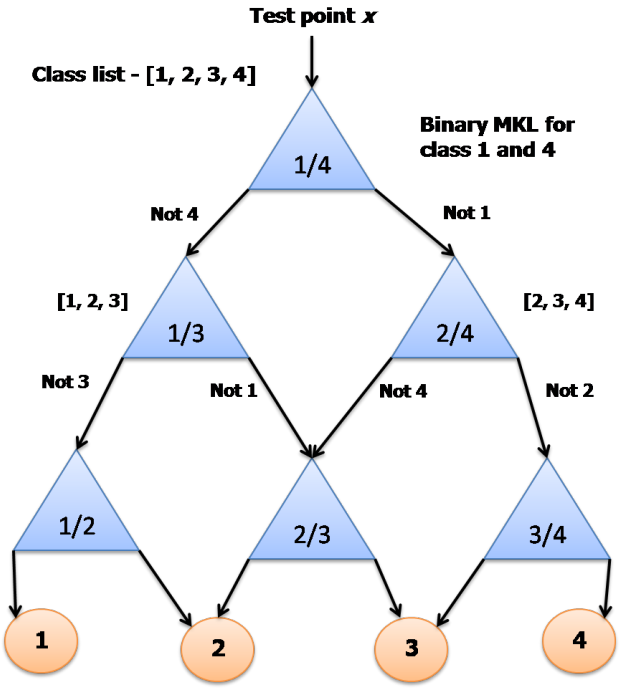


Figure 7.3: Concept based image annotation in 4 classes with binary MKL in DAG architecture

7.2.3 Experimental Results

The experimental evaluation of proposed annotation model is performed on posture image collection of two prominent Indian classical dance styles namely Odissi and Bharatnatyam[1]. The classical dance performances includes sequence of dancer's postures depicting different movements. We have prepared an image collection depicting various postures of these dances. In the posture image collection, the Bharatnatyam collection belongs to 11 categories and Odissi collection belongs to 12 categories.

Number of clusters i.e. vocabulary size is an independent parameter for indicator

[1]www.culturalindia.net/indian-dance/classical/index.html

vector computation. We have plotted the ROC (Receiver operating characteristics) curve with respect to different number of clusters. The best set of clustering parameters are used for subsequent experiments. The figure 7.4 shows ROC curve plotted for Bharatnatyam dance annotation model using local texture features. The KNN classifier is used as annotator. The curve is plotted for range of vocabulary size having 5 to 500 *visual words* while considering complete image set for clustering. Based on the ROC curve, the indicator vector computation using texture feature is performed with vocabulary size of 100 visual words. Similarly, the indicator vector computation for SIFT features is performed for 100 clusters. For Odissi dance images, the indicator vector using texture feature is computed with 100 clusters while complete dataset for clustering. Similarly, the indicator vector computation for SIFT features is performed for 50 clusters.

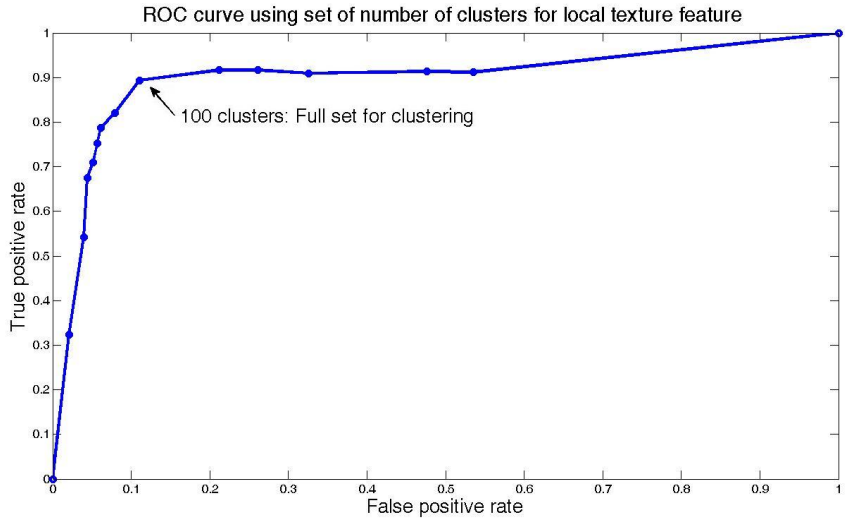


Figure 7.4: ROC curve Bharatnatyam dance posture annotation using local texture feature

First, posture concept annotation of images using different features are done with KNN and SVM classifier. In addition, the annotation results using concatenation of features are also presented. The annotation results for Odissi image collection is in Table 7.1, and for Bharatnatyam image collection is in Table 7.2. Here, LWV is the abbreviation for Local texture feature. SIFT and LWV contain significant amount of complementary information. For both the dance styles, the information is efficiently combined by the MKL framework shown by the improved classification results. The complementary nature of information existing in SIFT and local texture features give the best annotation results for both dance styles.

Table 7.1: Annotation accuracy for Odissi

Individual features			Concatenation of features	
	SIFT	LWV		SIFT–LWV
KNN	87.63	88.26	KNN	88.66
SVM	92.04	91.96	SVM	89.98

MKL based combination of features	
SIFT + LWV	93.84

Table 7.2: Annotation accuracy for Bharatnatyam

Individual features			Concatenation of features	
	SIFT	LWV		SIFT–LWV
KNN	79.46	83.20	KNN	89.00
SVM	81.22	84.18	SVM	90.24

MKL for single & combination of features		
SIFT	LWV	SIFT + LWV
87.14	89.42	94.81

7.3 MKL for LSCOM Concept Recognition

The experiments performed in Section 7.2.3 evaluate the DDAG architecture based multi-class MKL for domain specific image annotation. We can use the same framework for semantic concept based image annotation on subset of Columbia374 dataset. The dataset provides keyframes of TRECVID 2005 and 2006 dataset annotated using 374 concepts selected from LSCOM ontology [345]. A keyframe or still image represents the center of the shot segment extracted from the TRECVID videos. The features described in [345], viz. Edge direction histogram (EDH), Gabor filter response (GBR), and Grid color moment (GCM) have been used for feature space representation. The feature set corresponding to the selected concepts is divided in 10/90 proportion as testing and training set. For MKL based experiments, we have used set of linear and Gaussian kernels. The annotation results are presented in table 7.3. The non-linear SVM is trained with Gaussian kernel, and parameter selection is done by partitioning the training set in 30/70 proportion as validation and training set. The SVM based multi-class annotation is performed by DDAG architecture of binary SVMs. The original LSCOM annotations assign multiple annotations to many video shot. Therefore, we have considered only those keyframes which have uni-label annotation. The evaluation shows MKL based classification has improved the annotation accuracy by $0.82 \sim 1.54\%$ in comparison with SVM. The next evaluation using pair-wise combination of features showed that GBR and GCM achieved best accuracy (3.48% improvement with respect to individual best using MKL) as both the features represent complementary information. However, subsequent inclusion of

EDH increased the annotation accuracy marginally because of the overlapping nature of information between GBR and EDH.

Table 7.3: Image annotation accuracy using LSCOM concepts

SVM based annotation using individual features		
EDH	GBR	GCM
26.34	31.48	28.83

MKL based annotation using individual features		
EDH	GBR	GCM
27.88	32.43	29.65

MKL based annotation using combination of features			
EDH + GBR	EDH + GCM	GBR + GCM	EDH + GBR + GCM
34.64	30.77	35.91	35.96

7.4 MKL based Feature Combination for Concept driven Retrieval

The class labels associated with images have the notion of high level semantic concepts and represent significant amount of inherent semantics. The labels, therefore, provide efficient approach for concept based retrieval. Section 7.2 presented multiple feature based semantic concept learning for recognition using MKL. In the section 5.3, MKL based hashing is presented for feature based indexing. In the following discussion, we propose multiple feature based semantic concept learning using MKL for indexing. The proposed scheme therefore provides a novel concept based retrieval framework using multiple features. The experimental evaluation of the framework is performed on CIFAR-10 dataset [18]. The

dataset is subset of 80 million image dataset [14]. Considering the large scale of parent dataset, it covers wide range of variation. The image collection consists of 60000 colour images of size 32×32. The dataset has been primarily used for object recognition problem as each example is available with object level single label annotation. The annotations define concept classes, which have been used for optimization problem solution for MKL in retrieval framework. The collection is uniformly grouped in 10 categories and the dataset is partitioned as 50000 training and 10000 testing images. The feature descriptions and experimental settings are discussed as follows.

7.4.1 Image Feature Description

Since MKL based hashing is used for retrieval, set of feature descriptors representing diverse image characteristics have been applied. The first feature: GIST, computes global image representation by describing complete image scene as single entity [229]. The descriptor is computed by extracting the statistics of low level image features. The image is first decomposed by multi-scale oriented filters. The output of these filter responses are averaged on 4×4 grid. Therefore, for 8 orientations and 3 scales, each image is represented by 384-d vector.

The Local texture feature described in Section 7.2.1 is used as second descriptor. The neighbourhood window of 7×7 is selected for local wavelet response computation as example images are of small size and the window overlaps region of the image. The selection of optimum dictionary size for bag-of-words computation is done by evaluating the classification accuracy by cross validation over sampled image set from training

set. First the bag-of-words representation for training images is generated for different dictionary size, and then 5-fold cross validation based classification accuracy over 2500 randomly selected examples is computed using KNN classifier. Five nearest neighbours are considered for majority voting. Figure 7.5 shows classification performance. Based on the performance, images from the dataset are represented by bag-of-words representation computed with dictionary size of 50. Therefore, each image is represented by 50-d vector.

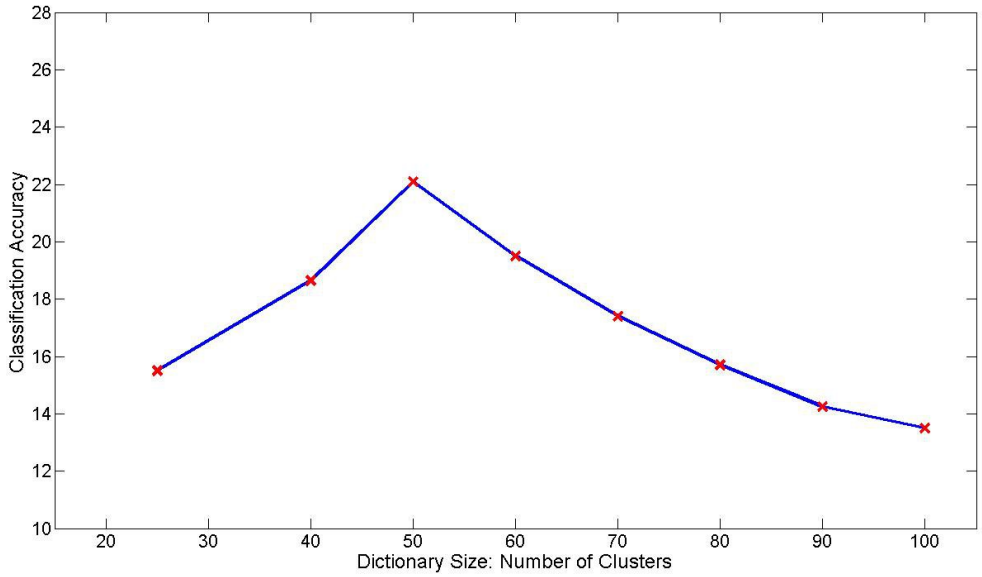


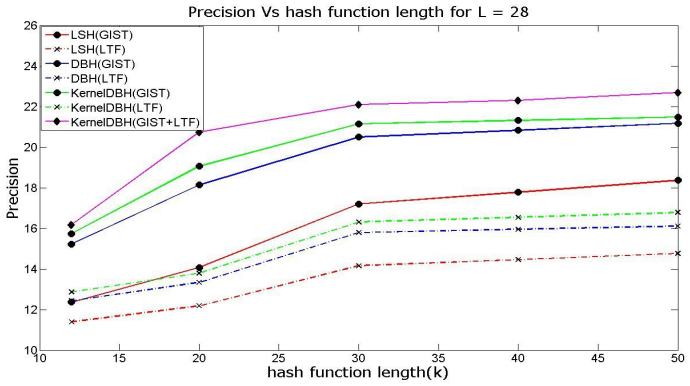
Figure 7.5: 5-fold cross validation classification accuracy for different dictionary size

7.4.2 MKL Details and Results

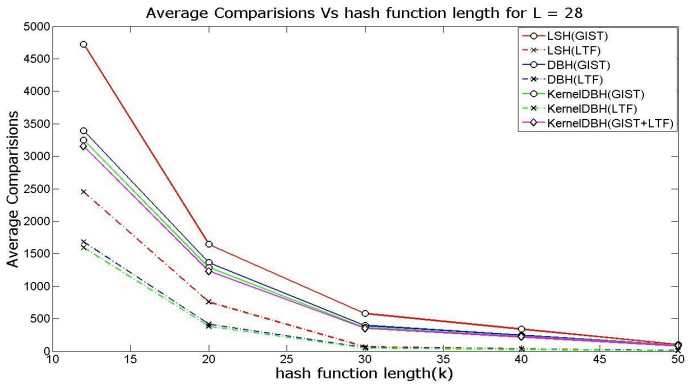
The GA parameters for MKL are selected same as discussed in section 5.3.3. The parameter setting is represented as

- Initial population consisting of 40 strings.
- Crossover and mutation probability selection as 0.8 and 0.05.
- Kernel weight w_i encoding by 5-bit binary string.

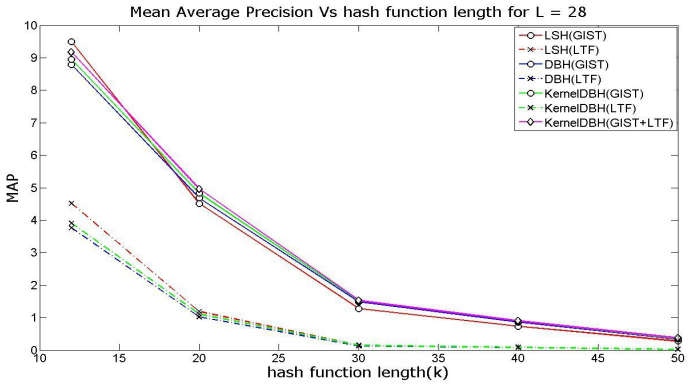
The GA iteration is simulated for 100 generations. The GA fitness evaluation is performed by computing MAP for the evaluation query set X_v. The validation query set X_v consists of 1000 images selected by stratified sampling from the training images. The hash function generation is done by selecting 1000 images by stratified sampling from the remaining training images (49000 images). The base kernel set for GIST features includes a linear and set of Gaussian kernels with variance: $\{0.1, 0.5, 1, 2, 5, 10\}$. The set of base kernels for texture feature included a set of Gaussian kernels with variance: $\{0.1, 0.2, 0.5, 1, 2, 5\}$. For learning the optimal combination of both features for indexing, the base kernel set is formed by ORing individual base kernels. The generation of $\mathbb{H}_{\mathbb{DBH}}$ & $\mathbb{H}_{\mathbb{KDBH}}$ for performance evaluation is done by stratified sampling of 1000 images from complete training image set. The experimental results are presented in figure 7.6 and 7.7 (Texture feature is abbreviated as LTF). The retrieval results with random hyperplane based LSH are also presented. Again the KernelDBH achieved better performance than the DBH and LSH. In particular, significant improvement in precision values is obtained by learning based combination of GIST and texture descriptor $\{(3.17 \sim 17.66\%)$ over baseline KernelDBH results$\}$. Individually, the retrieval performance by GIST based representation is better than the texture descriptor based representation. The wavelet function families demonstrate good localization property in both frequency and time; therefore, efficiently extract the local



(a) Precision Values

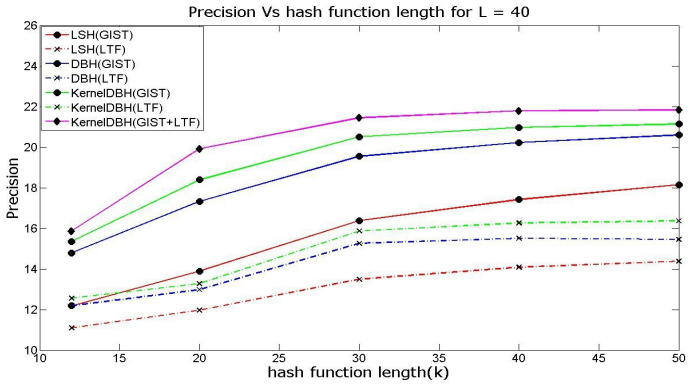


(b) Average Comparisons

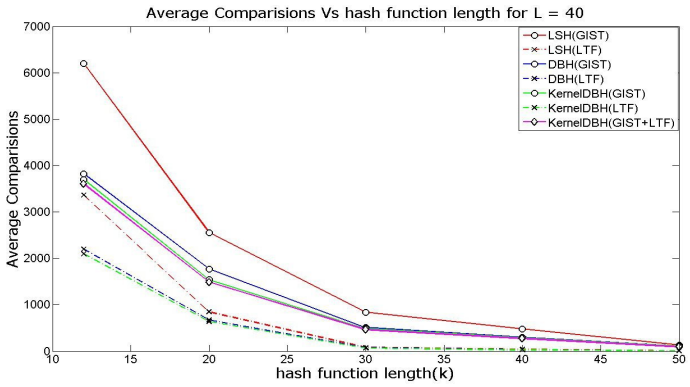


(c) Mean Average Precisions

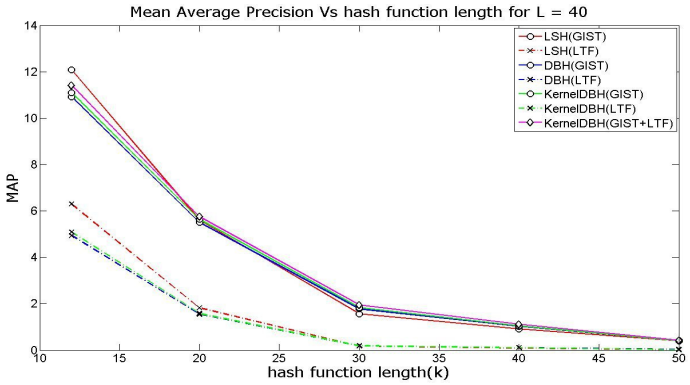
Figure 7.6: Results with CIFAR-10 dataset: $L = 28$



(a) Precision Values



(b) Average Comparisons



(c) Mean Average Precisions

Figure 7.7: Results with CIFAR-10 dataset: $L = 40$

image features. However, the small size of training images generate sparse distribution of key points resulting the texture descriptor with insufficient discriminative power. The results show that proposed MKL for indexing carefully combines both the descriptors such that Kernel based DBH efficiently exploits the benefit of complementary informations of both the features to improve its grouping performance in hash space.

7.5 Multi-modal Concept linkage using Conditioned Topic Modelling

In practice, various multimedia applications deal with data having information in multiple modes. In chapter 6, we presented MKL based methods for text document retrieval having multi-modal information. The performance of concept based retrieval system significantly depends on the understanding of the semantics of content. The multi-modal information are mostly complementary in nature. In case of availability of such information, semantic understanding of the content improved by exploiting the correspondence information. In this context, concept level modality fusion provides logical approach for multi-modal content retrieval. The following discussion presents framework for learning the concept level semantic correlation between document modalities by applying combination of generative and discriminative modelling. As discussed above, the availability of concept lexicon is not always guaranteed. Here, the topic models for text modelling provide an efficient solution [67, 135, 31]. These models explore the underlying semantic structure of a document collection and discover the semantic grouping of existing terms. In recent works, topic

models have been successfully extended for image analysis. In this work, we begin with cross-modal content retrieval framework proposed in [148] and propose conditioned topic learning in LDA based indexing to learn the latent topic assignment aligned with their contextual grouping.

7.5.1 Conditioned Topic Learning for Multi-modal Retrieval

The framework presents generative-discriminative modelling approach for indexing the multi-modal documents. The model learns the multi-modal concepts from the document collection by defining conditioned topic modelling method. The proposed concept applies LDA based generative model for semantic indexing of documents.

The Topic Learning Model

The proposed model generalizes the LDA modelling over multi-modal document collection. The model here learns the relationship between latent topics generated from different modalities. Considering $\mathcal{D}$ as the multi-modal document collection. The indexing model for the document sets assumes that each document contains a modality index y_k, i.e. the set of documents is represented as $\mathcal{D}$ = $\{(y_1, \mathbf{w}_1)$,$(y_2, \mathbf{w}_2)$, ... ,$(y_{|D|}, \mathbf{w}_{|D|})\}$. The set $\{1, ..., V_M\}$ contains the vocabulary size of different modalities and each word w_{kn} for $(1 < n < N_k)$ assumes a discrete value from this set. Evidently, the definition of set $\mathcal{D}$ relaxes the condition of one-to-one correspondence between documents from different modalities. The information of context level relationship between the documents is expressed by a similarity graph $\mathcal{G} = (\mathcal{D}, \mathcal{E})$ which is computed at the topic level. The

multi-modal document random field model presented in [148] defines the similarity graph by considering the pair-wise potential. The present method expresses the contextual relationship between documents by applying unary and pair-wise potentials. The objective is the incorporation of broad level contextual information for the latent topic learning over multi-modal document collection. The potential function for each node is defined as

$$E(\theta_i) = exp(-\lambda_1 f_1(\theta_i) + -\lambda_2 f_2(\theta_i, \theta_j)) \tag{7.5.1}$$

In this sense, the graph represents a conditional random fields over the document collection. Here unary potentials are posterior probability estimate by a learned classifier using the individual topic distribution corresponding to documents from different modalities. The pair-wise potential estimates are defined as the symmetric KL divergence between topic distributions.

The generative procedure of the proposed topic modelling follows the conventional LDA generative procedure, which first samples $\boldsymbol{\theta}_k$ for k^{th} document, and subsequently samples words for k^{th} document with respect to $\boldsymbol{\theta}_k$. However, the topic distribution in present scenario also considers the contextual similarity information of documents. The similarity graph G expresses the document relationship as well as contextual category information of different documents. The generative process is summarized as follows:

- Sample the word distribution for m^{th} modality as $\boldsymbol{\phi}_m \sim Dirichlet(\boldsymbol{\phi}|\beta_m)$

- Sample the document-topic distribution $\boldsymbol{\theta}_{1,...,|D|}$ as

$$p(\boldsymbol{\theta}_{1,...,|D|}|\alpha, \mathcal{G}) = \frac{1}{Z} exp\{-\lambda_1 \sum_{k=1,...|D|} f_1(\boldsymbol{\theta}_k) - \lambda_2 \sum_{k,j \in \mathcal{E}} f_2(\boldsymbol{\theta}_k, \boldsymbol{\theta}_j)\} \times \prod_{k=1,...|D|} Dirichlet(\boldsymbol{\theta}_k|\alpha) \tag{7.5.2}$$

- Sample topic z_{kn} for word w_{kn} from $Multinomial(z|\boldsymbol{\theta})$
- Sample word w_{kn} from $Multinomial(w|\boldsymbol{\phi}_{mz_{dn}})$

The joint probability for document collection with incorporation of similarity as well as contextual category information is expressed as

$$\begin{aligned} &p(\mathcal{D}, \boldsymbol{\theta}_{1,...,|D|}, \boldsymbol{z}_{1,...,|D|}, \boldsymbol{\phi}|\alpha, \beta_{1,...,|M|}, \mathcal{G}) \\ &= \frac{1}{Z} \prod_{m=1}^{M} \prod_{k=1}^{K} Dir(\phi_{mk}|\beta_k|\alpha)[exp\{-\lambda_1 \sum_{k=1,...|D|} f_1(\boldsymbol{\theta}_k) - \lambda_2 \sum_{k,j \in \mathcal{E}} f_2(\boldsymbol{\theta}_k, \boldsymbol{\theta}_j)\}] \\ &\quad \times \prod_{k=1,...|D|} Dir(\boldsymbol{\theta}_k|\alpha)(\prod_{n=1}^{N_k} Mult(z_{kn}|\theta_k) Mult(w_{kn}|\phi_{y_k} z_{kn})) \end{aligned} \tag{7.5.3}$$

Inferencing and Parameter Estimation

The form of the equation (7.5.3) shows that exact inferencing is not possible because of the intractable form of the likelihood equation due to coupling between $\boldsymbol{\theta}$ across the document collection. Following the approach presented in [30], we adopt empirical conditioned topic learning model where context level similarity enforced by conditional random field enforced by an empirical topic distribution $\hat{\boldsymbol{\theta}}_d$ for document and compute the unary and pair-wise potentials using these distributions. The empirical topic distribution $\hat{\boldsymbol{\theta}}_d$ is

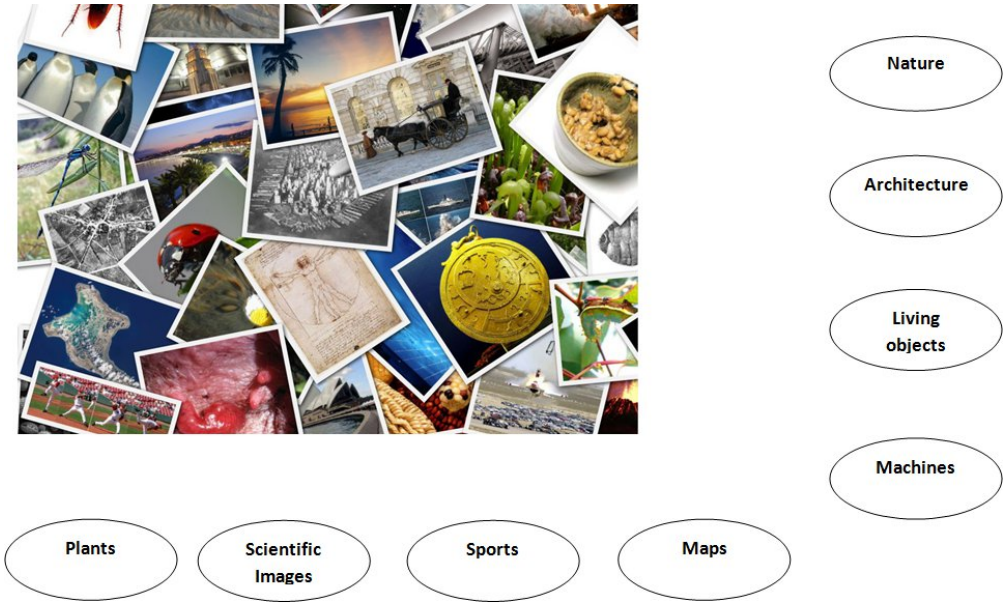


Figure 7.8: Sample images and corresponding subject based categories

computed as

$$\hat{\theta}_{kl} = \frac{n_{kl} + \alpha}{\sum_{l=1}^{T} n_{kl} + T\alpha}$$

Here T is the number of topics. The parameter $T\alpha$ introduces smoothness in the distribution. We apply Gibbs sampling for parameter estimation in the proposed topic learning model [128]. Gibbs sampling is a special case of Markov-chain monte carlo (MCMC) simulation which provides simple algorithms for approximate inferencing in high-dimensional models such as LDA. The process applies latent-variable method of MCMC simulation where the parameters $\boldsymbol{\theta}$, and $\boldsymbol{\phi}$ are integrated out as they are interpreted as the statistics of associations between the observed w_{kn} and the corresponding z_{kn}, the state variable of the Markov chain. In this context, the strategy of integrating out some of the parameters for inferencing is generally referred collapsed Gibbs sampling. In

the collapsed Gibbs sampling, $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are integrated out and sampling is performed for $\mathbf{z}$. For k^{th} document from m^{th} modality, the probability of l^{th} topic assignment to word is computed as:

$$p(z = l|\mathcal{D}, \mathbf{x_w}, \alpha, \beta) \propto$$
$$\frac{n_{kl} + \alpha}{\sum_{l=1}^{T} n_{kl} + T\alpha} \times \frac{n_{lw} + \beta_m}{\sum_{w=1}^{V_M} n_{lw} + V_M\beta_m} \times$$
$$\{ \prod_{k',(k,k')\in\mathcal{E}} exp\{-\lambda_1\{f_1(\boldsymbol{\theta}_{k,-z}) - f_1(\boldsymbol{\theta}_{k,z=l})\} - \lambda_2\{f_2(\boldsymbol{\theta}_{k,-z}, \boldsymbol{\theta}_{k'}) - f_2(\boldsymbol{\theta}_{k,z=l}, \boldsymbol{\theta}_{k'})\}\} \quad (7.5.4)$$

$\theta_{k,z=l}$ is the empirical topic distribution for k^{th} document with w assigned the topic l, and $\theta_{k,-z}$ is the topic distribution for k^{th} document without considering word w. For each iteration of Gibbs sampling based parameter estimation, given S subject categories in the document collection, we compute S probabilities: $P(1|\hat{\boldsymbol{\theta}}_1), \ldots, P(S|\hat{\boldsymbol{\theta}}_d)$. Here $P(l|\hat{\boldsymbol{\theta}}_j)$ denotes the probability of $\hat{\boldsymbol{\theta}}$ belonging to subject l. The values here are the notion of the top-down uncertainty of document content [313]. The objective here is to align the final topic assignment to most confident subject category. The unary potential represented by f_1 is estimated as $\sum_i^S logP(i|\hat{\boldsymbol{\theta}}_d)$.

7.5.2 Experimental Results and Discussion

The proposed concept is experimentally validated on the dataset discussed in [148]. The dataset contains 2600 images of variety of subjects and corresponding textual description describing its information content. Sample images are shown in the figure 7.9. A subset of 676 examples images is randomly selected and the subset is annotated in eight categories

based on the subjective analysis of the image content and corresponding description. The categories basically represent the context groups defining a major subject. The details of the categories are as follows.

Table 7.4: Description of different subject categories

Broad level category	covered subjects
Nature	Landscape, sky, oceans, mountains
Architecture	Buildings, bridges
Living objects	Humans, birds, animals, insects
Machines	cars, planes, bikes
Maps	hand-drawn and satellite images, portraits, paintings
Sports	All the sports related images
Scientific Images	Constellations, scientific equipments and description charts
Plants	Grains, vegetables and flowers

The text available with image present loose description of the information contained. For the text cleaning purpose, words having less than 4 characters, and words having less than 5 occurrence in the dataset are filtered out at pre-processing stage. The unary potential for the empirical topic distribution estimation is computed by Bayesian probability estimate as discussed in section 7.5.1. The probability estimate is computed by Relevance vector classifier learned on the image annotated with the above defined subject categories. The image feature representation based on bag-of-words model is computed by the SIFT descriptors. The bag-of-words based representation is computed with code-book having 1000 visual words. The methodology discussed in [148] is applied for evaluation. The performance is evaluated by retrieving the relevant images corresponding for given text

queries. For text query $\mathbf{w} = \{w_1; w_2; .....; w_n\}$, the dataset images are ranked using

$$p(\mathbf{w}|\boldsymbol{\theta}_i) = \prod_{j=1}^{n} p(w_j|\boldsymbol{\theta}_i)$$

The $\boldsymbol{\theta}_i$ represents topic distribution for the i^{th} image of the collection. The marginal probability estimate $p(w_j|\boldsymbol{\theta}_i)$ for all the text terms are computed while training. Using a synthetic query set, the presented approach achieved 50.76% accuracy in the first 20% of ranked list of results. For the same query set, the MDRF method discussed in [148] retrieved 48.04% accuracy. The result establish that conditioned topic modelling efficiently exploits the dependency between multi-modal features and labels resulting more accurate semantic grouping.

7.6 Multi-modal Concept Recognition

In section 7.5, combination of generative and discriminative modelling is explored for learning the multi-modal topic distribution for retrieval. In the subsequent discussion, we use the generative and discriminative modelling in succession for recognition problem by learning multi-modal concepts. The details of the framework is shown with respect to event detection application in broadcasted sport videos. Also, the applicability of framework is shown for concept based labelling of document images having multi-modal information. Event detection has acquired considerable research interest for the development of video summarization and surveillance applications. The proposed recognition framework applies the Latent Dirichlet Allocation (LDA) based topic modelling for extracting the semantic primitives defining different sport events. We capture the temporal and spatial combination

Abbey of Senanque, located in France, Provence, Vaucluse, Gordes village. An abbey is a Christian monastery or convent, under the government of an Abbot or an Abbess, who serve as the spiritual father or mother of the community.

The Loch Ard Gorge, found in Port Campbell National Park, Victoria, Australia, is named after the clipper ship Loch Ard, which ran aground on nearby Muttonbird Island on 1 June 1878 approaching the end of a three-month journey from England to Melbourne. Shown here is a panorama of 4 segments taken from the cliffs looking down towards Loch Ard Gorge.

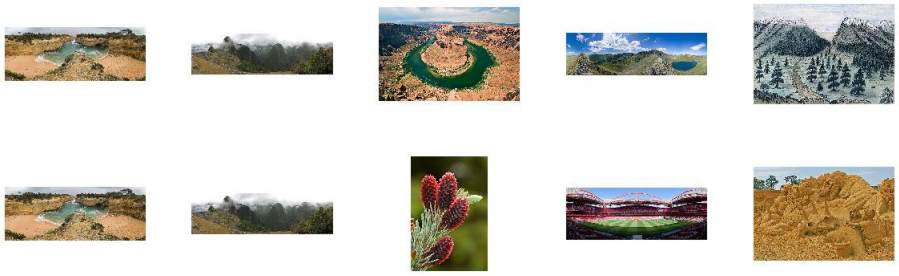


Figure 7.9: Sample retrieval results corresponding to the textual description in the left: The top row shows retrieved image from the proposed method and bottom row shows the images retrieved by the method presented in [148]

of these semantic primitives by Conditional Random Fields (CRFs) based probabilistic graphical model to identify various sport events.

Automated video event detection primarily consists of two sub-problems, 1) event representation and 2) modelling. Events are basically defined as sequence of activities

performed by different objects. Here, the activity identification is mostly restrained to a particular sport because of its unique set of characteristic activities. The existing works in this direction have developed various sport specific feature detectors by utilizing low-level attributes e.g. color, edge and transforms. However, such schemes have limited generalizing capability across different sports. Additionally, the gap between low-level features and the high level semantic description of an activity also affects the overall performance of the system. In this work, LDA is applied to extract the characteristic features from video segments. The approach provides a robust method to identify characteristic features by semantic space grouping of low-level features in video segment. The low-level features are extracted in the form of spatio-temporal descriptors from video stream and Mel-frequency cepstral coefficients (MFCC) [212] from Audio stream. The topic modelling performs latent space grouping of these features exploring their contextual relationship. Topics in the present context are the instances of semantic activities defining various events.

The primitives represented by video and audio topics define the set of semantic activities for sport events. However, these primitives do not preserve temporal sequence information characterizing sport events. We apply CRFs based graphical model for preserving the sequential dependencies between topics for event prediction.

7.6.1 Proposed Event Detection Framework

The proposed event detection framework addresses the above discussed sub problems by defining a hierarchical scheme (Refer figure 7.10). The initial step applies generative modelling to extract the semantic characteristic features from the video and audio streams

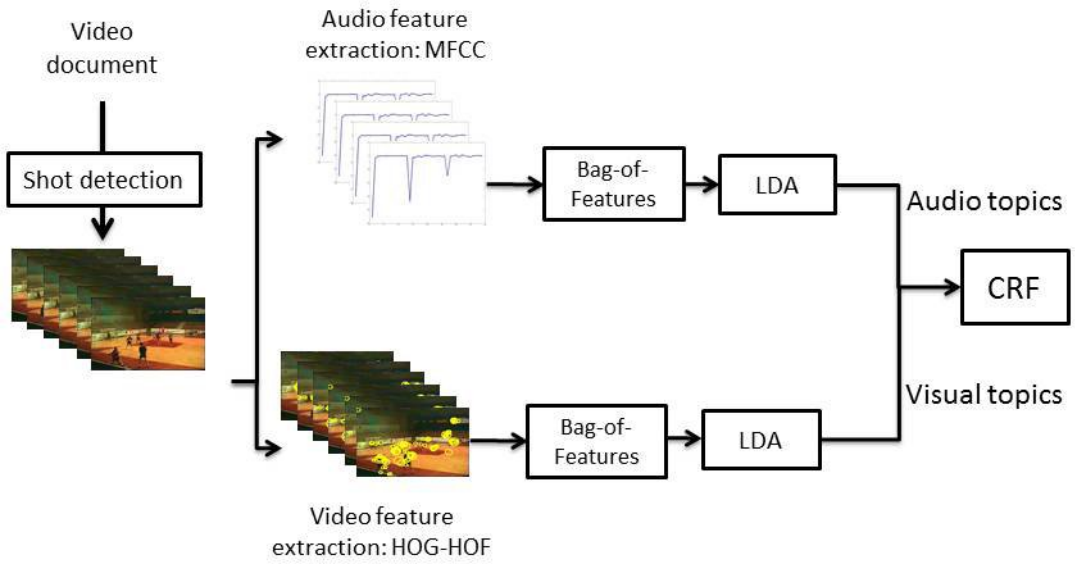


Figure 7.10: Event Detection by topic based CRFs

corresponding to video segment. The second step performs discriminative modelling over the training characteristic feature set for event modelling.

Characteristic Feature Extraction from Video Segments Using LDA

The video data representation forms an important step for event interpretation. We extract local spatio-temporal descriptors following Laptev et al. [170] as low level features. The descriptor computation is initiated by identifying the interest points in frame sequence by multi-scale analysis, as the points having large variations along the spatial and temporal directions. Two set of descriptors, histogram of gradients (HOG) and histogram of optic flows (HOF) are computed in the space-time neighbourhood of interest points to characterize local motion and appearance. Using the set of local descriptors, bag-of-video-features (BoVF) is computed for a video segment. The clustering of spatio-temporal codebook

generation is done by k-means algorithm. The BoVF represents the histogram of spatio-temporal descriptors over space and time by Euclidean distance based assignment of local descriptors to nearest word in the codebook. Given the BoVF for video segments, the next task is to extract the characteristic features. We perform LDA modelling to identify the hidden concepts in video segments as the representative characteristic features. As the concepts define the context based relationship between spatio-temporal descriptors, they equivalently represent the semantic primitives as the notion of latent characteristic features. The learning based approach for feature detection therefore enhances generalization of the proposed framework across sport categories.

The inferencing step of LDA modelling assigns a topic distribution for a video segment which is subsequently applied for event modelling. Following similar approach, the topic distribution for audio streams associated with the video segments are extracted. Bag-of-audio-features from audio streams is computed by segmenting the stream in smaller segments of 40ms (assuming 25 frame per second). Each smaller segment is an audio term represented by MFCC features. LDA modelling for topic extraction follows the similar approach as discussed earlier.

Event Modelling Using CRFs

Topic distribution for video segments represent the mixture of semantic activities defining different events. The latent topics in the video segments are self evolving in nature due to unsupervised learning. However, we need to learn the alignment of these topics for different event categories by performing supervised learning. Also, the characteristic features for

event detection are interrelated and exhibit significant corresponding information. For example, presence of goal post in a soccer video segment implies the start, break, defence or offence event. However, all these events require presence of ball near goalpost except during break. In such context, supervised learning using probabilistic graph model can learn the topic sequence alignment and their contextual dependencies. We apply CRFs based probabilistic graphical model for modelling the sequential information [164]. CRF models the observed variable based on the current as well past states, thus efficiently captures the long-range dependencies in feature sequences.

Individual information from video and audio provide complementary evidence for the occurrence of an event. However, a segment having far view of crowd will have similar visual description for a non interesting segment and segment depicting 'mexican wave', in this case audio impulse provides the discriminatory confidence. Therefore we augment both the information to improve the accuracy and robustness of our system. The augmentation is performed by concatenating the video and audio topics. The combination of video and audio topics define nodes of the graph G. The modelling of these topics develop into a probabilistic temporal model which can identify significant events corresponding to the video segment.

7.6.2 Experimental Results

The experimental evaluation of the proposed framework is performed on two different sport videos. First evaluation is performed on the Handball video available at [60]. The second evaluation is performed on recorded soccer video downloaded from the Internet.

The experimental details and results are discussed as follows.

Experiment with Handball Video

The duration of European handball video available at [60] is 10 minutes and the video recording is performed by hand-held camera. The frame rate of the video stream is 25 and resolution is of 384×288. After manual annotation, 67 segments have been identified based on 9 different team activities. Low-level visual feature extraction is performed as discussed in the section 7.6.1. Corresponding to video segments, sets of local descriptors as HOG and HOF features are computed. Following the descriptor parameters suggested by authors in [170], HOG and HOF are computed as vectors having 72 and 90 elements. The combination of descriptors by concatenation is applied following the results shown in [319]. The codebook size is an important performance parameter in the present framework as it identifies the set of video and audio patterns reflecting the video contents. Therefore, in this work, we evaluate the proposed framework for different codebook sizes to identify the suitable codebook sizes.

The LDA based topic extraction assigns video and audio topic distribution for a segment. In general, visual content exhibit more variance than audio in sport videos leading the selection of more number of video topics than audio for LDA modelling. The evaluation is performed for set of video and audio topics as $\{10, 20, 30\}$ and $\{5, 8, 10\}$. CRFs based modelling for event detection is performed by combining the video and audio topics. However, the topics with weak strength are required to be filtered for robust detection. Therefore, we filter the insignificant topics from video and audio topic dis-

tribution by applying mean based threshold before combination. The results discussed here are simulated by computing the threshold as $\mu + 0.1 \times \sigma$, where μ, σ are the mean and standard deviation of probability distribution. Figure 7.11 shows results using visual

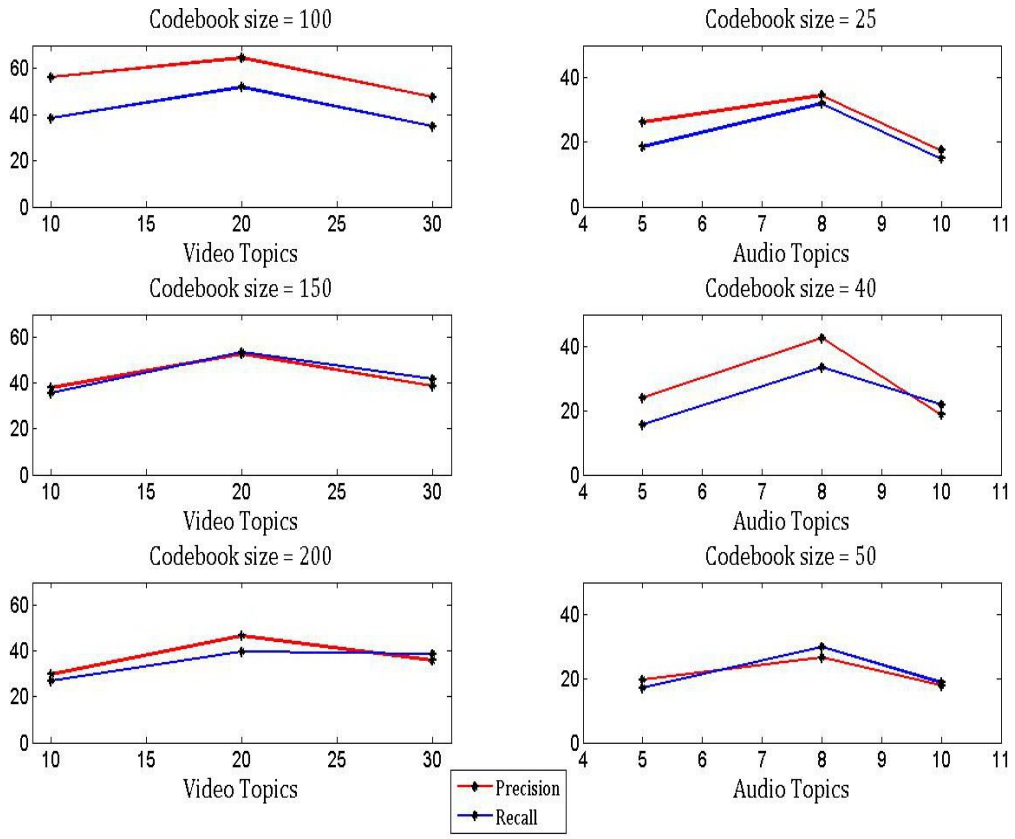


Figure 7.11: Detection using individual modality

and audio information for event detection in independent fashion. The results are average of five iterations performed after 10/90 stratified splitting of all samples for testing and training. It is clear that for visual and audio characteristics features extraction for 20 and 8 topics give better results. However the performance also depends on codebook size, we combined both modalities for 20 video and 8 audio topics for different codebook size.

Table 7.5 presents three best results from different combinations of modalities. The results show significant improvement in prediction performance by combination of topics. Nevertheless the experiment shows that maximum likelihood based solution may achieve comparable results from individually sub-optimal results.

Table 7.5: Results on Handball video

Video Paras.		Audio Paras.		Precision	Recall
Words	Topics	Words	Topics		
100	20	25	8	68.05	56.12
150	20	40	8	62.15	83.88
100	20	40	8	75.66	74.10

Evaluation on Soccer Video

The soccer video stream used for second evaluation of the proposed framework is of 35 minutes duration. The frame rate of the video stream is 25 and resolution is of 480×360. The video is segmented manually, into smaller clips of five second duration. Manual annotation of total segments is done in two broad categories i.e. {interesting and non-interesting}. The interesting video segments correspond to events comprising goal attempts, save, corner and penalty kicks and interruptions due to injury or misconducts. Remaining segments are identified as non-interesting. Based on these events, the annotation process identified 77 interesting and 343 non-interesting segments in the video stream.

Low-level feature extraction from video segments is done as discussed in the section 7.6.1 with original parameter setting as discussed in section 7.6.2. The event detection framework is evaluated for the set of video and audio topics as $\{20, 35, 50\}$ and

$\{10, 15, 20, 30\}$. For different codebook sizes ($\{100, 150, 250\}$ and $\{25, 50, 75\}$ for video and audio respectively), we observed 20 video and 10 topics achieved better results. For each codebook size best result and corresponding number of topics is presented are table 7.6. For all the codebook sizes 20 video and 15 audio topics achieved better results. The results are justifiable as the audio information from a soccer game is much more dense. Also in this case, the independent modality based detection results are comparable. Three best

Table 7.6: Results on Soccer video

Using Video				Using Audio			
Words	Topics	Precision	Recall	Words	Topics	Precision	Recall
100	20	70.21	83.12	25	15	53.96	48.14
150	20	76.78	82.68	50	15	68.34	72.10
200	20	54.94	43.83	75	15	43.08	53.68

Video Paras.		Audio Paras.		Precision	Recall
Words	Topics	Words	Topics		
100	20	25	15	78.35	86.10
150	20	25	15	81.40	92.86
150	20	50	15	89.18	92.78

results by CRF modelling using combination of information is presented in table 7.6. These results are computed as average of five iterations performed by 20/80 stratified splitting of samples for testing and training. The results again establish the efficacy of the proposed framework by effective utilization of video and audio information. Also the video and audio codebook of 100, and 150 terms give better results. In general, the patterns from video segment extracted for smaller codebook are not sufficiently discriminative. Additionally, large codebook performs noisy allocation of terms to different codebook words. However,

the understanding of the effect of codebook size after combination of informations require more experiments on different sport videos.

7.6.3 Concept Recognition of Multi-modal Document Images

The video event detection framework discussed in section 7.6.1 presents a generalized multi-modal concept based recognition framework. The LDA based modelling for topic extraction presents a generalized framework for exploring the latent semantics of multi-modal documents. The application of CRF for learning the context between topic sequences from different modalities extends the framework's applicability for semantic labelling of document images having multi-modal images. Figure 7.12 shows the block diagram of application framework.

The framework is evaluated on sample document image collection from English mag-

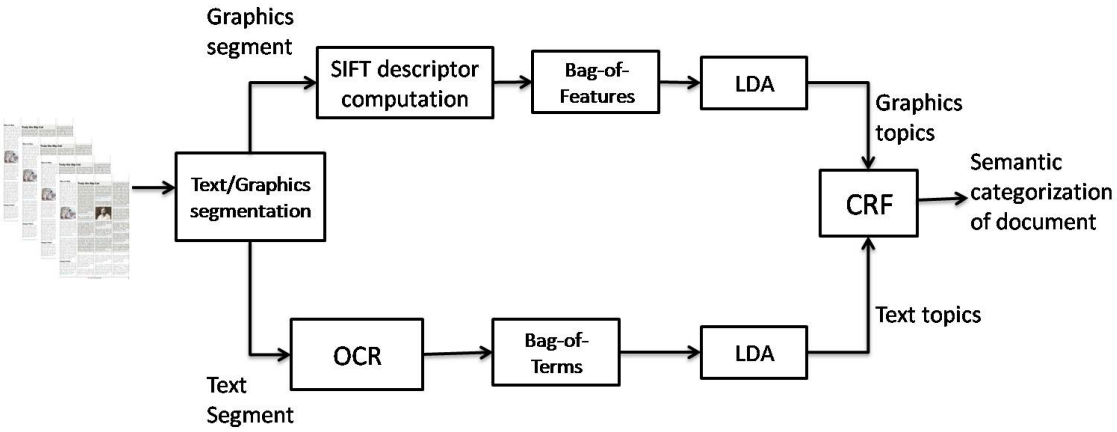


Figure 7.12: Semantic labelling of multi-modal document images by topic based CRFs

azines. The collection consists of 211 pages having articles from four broad categories

defined as sports, politics, entertainment and commerce. The graphics regions from document images are segmented following the framework discussed in section 6.2. The graphics segments less than $1/50^{th}$ of document size are filtered out. The text image segments have been OCR'ed by using Abbyfine Reader [303]. Subsequently, text terms having less than 4 characters are filtered out. The SIFT features are extracted from segmented graphics regions which are converted to bag-of-words representation for vocabulary size of 50 visual words. Therefore, each segment is represented by 50-d vector. The initial experiment is performed as semantic labelling using individual modalities. The evaluation is performed in 5-fold cross-validation setting. For both type of information, the number of topics are varied as $\{10, 25, 50, 100\}$. The results presented in table 7.7 show that for the experimen-

Table 7.7: Results of semantic classification of document images

Using Graphics		
Topics	Precision	Recall
10	12.35	17.83
25	28.94	31.65
50	22.12	25.54
100	17.79	16.28

Using Text		
Topics	Precision	Recall
10	29.66	36.78
25	44.85	37.91
50	52.30	60.10
100	34.51	30.19

Combination of text and graphics			
Topics		Precision	Recall
Text	Graphics		
25	25	41.85	39.91
50	25	63.23	66.85
50	50	56.72	62.41

tal document collection, latent topics extracted from text segments are more discriminatory and conclusive. Using individual information, graphics and text based recognition showed

better results for 25 and 50 topics. In the subsequent experiment, the text and graphics are combined using different combination of number of topics. The results in table 7.7 show that concept level fusion of multi-modal information improved the classification precision by 10.93% and recall by 5.75% with respect to individual best. Identification of semantic concepts for specific domain is a challenging task.

7.7 Conclusions

The chapter presented semantic concept based multimedia content analysis methods. First, methods to learn semantic concepts using the multiple feature based representation are presented. The methods apply the MKL based feature combination, which have been applied for posture based annotation of dance images and concept based natural image retrieval. Novel feature representation is presented which represents the local texture property of images in bag-of-words model. The efficacy of the feature is shown for concept based image annotation and retrieval experiments.

Novel methods for multi-modal concept based document recognition and retrieval framework are presented. We have proposed conditional topic learning for multi-modal indexing and retrieval. The probabilistic model presented here combines the generative and discriminating learning for multi-modal retrieval. The results showed that multi-modal topic learning conditioned over the generalized subject categories improved the retrieval performance in comparison with the existing results. Finally, multi-modal concept based document recognition framework is presented, which exploits generative modelling for

extracting the semantic primitives describing the document characteristics. The CRF based discriminative model is applied to learn the the sequence of semantic primitives for different document categories. The framework presents a generalized approach which can be applied for semantic categorization of different types of multi-modal documents. The primary evaluation for event detection in sport videos establish the efficacy of the framework. Additionally, the applicability of the framework is demonstrated for semantic categorization of multi-modal document images.

Chapter 8

Conclusions

In this thesis, machine learning based methods for multimedia document retrieval and recognition have been presented. Specifically, the following problems have been addressed using machine learning.

→ Feature representation for binary patterns.

→ Indexing of documents having text in imaged form.

→ Feature combination for multimedia document recognition and retrieval.

→ Identification of text and graphics regions from document images.

→ Script identification from documents having intermixed multi-lingual text.

→ Multi-modal retrieval of multi-modal multimedia documents.

→ Semantic concept learning using multiple features.

→ Understanding of multi-modal concepts for recognition and retrieval.

The above discussed problems have been studied for different type of multimedia documents. First three problems have been studied in the context of multimedia documents having image form of textual information. Multiple feature based classification framework is proposed for the recognition of image based text documents. Further, learning based frameworks are presented for indexing and retrieval of document images using single and multiple feature word image representation. The remaining problems address the class of multimedia documents having co-existing and associated multi-modal information. Methods to integrate multi-modal information from document images including script identification are presented. Finally, frameworks for semantic concept learning, and correlation between multi-modal concepts are presented for multi-modal retrieval and recognition.

8.1 Summary of the Contributions

The summary of main contributions of the presented work is as follows:

* Shape based feature representation is presented for binary patterns such as characters, symbols and words. The feature provides object description by constant dimensional vector directly applicable for similarity matching and analytical model learning. Binary pattern recognition framework is presented using the single and multiple feature representations. The framework presents application of binary multiple kernel learning (MKL) in decision directed acyclic graph. The shape based feature with set of different features is applied on the framework for recognition of character

and symbol primitives of Indian scripts. The evaluation is also demonstrated for recognition of MPEG graphic symbols.

* Learning based framework for multimedia document indexing and retrieval using single and multiple feature definitions. Word based document image indexing framework is presented by the application of distance based hashing (DBH). Clustering based pivot object selection, and multi-probe hashing framework for DBH functions are presented. The efficacy of indexing framework is shown with shape based, and string like word representations. The kernel space extension of distance based hashing is developed. Using the extension, the MKL formulation for retrieval is formulated. The evaluation of MKL is presented on handwritten digit images. Using the MKL formulation, word based document image indexing and retrieval framework using multiple feature representation is presented. The extensive evaluation of the presented frameworks is performed on document image collection of Devanagari, Bengali and English script.

* Document image segmentation framework for identifying the text and graphics regions in documents having complex layout and overlapped modalities. Using the information of different regions, multi-modal document retrieval framework is defined by multiple kernel learning formulation for retrieval. Robust and fast script identification framework is defined for bi-lingual documents having mixed scripts. The framework for latent Dirichlet allocation (LDA) retrieval of text documents having recognition errors is presented. The framework presents method to incorporate

the recognizer's characteristics in indexing process for robust and accurate retrieval in case of recognition inaccuracies. Subsequently, the latent topics are applied for multi-modal indexing with corresponding document images for improved retrieval performance.

* Framework for semantic concept learning from multimedia documents using multiple feature representations is presented. The use of multiple kernel learning based application of multiple features for concept based image retrieval is demonstrated. Multi-modal document retrieval framework is defined by learning the cross-modal correlation across modalities using conditioned topic learning by combining generative and discriminative modelling. The combination of generative and discriminative modelling is subsequently explored for semantic concept recognition from multi-modal documents. The application of the framework is described for event detection application from sport videos, and semantic categorization of document images having text, and images.

8.2 Scope of Future Work

In this thesis, the problem of application of multiple features, and multi-modal information fusion for multimedia document retrieval is studied. In addition to experimental evaluation and validation, some interesting directions for extensions of presented concepts are as follows:

- Exploration of novel MKL formulation for multi-class problem addressing the re-

quirement of large number of examples as well as categories. In particular, defining novel mathematical formulation for MKL for the class of large-scale applications.

- The extension and evaluation of DBH for accommodating incremental updation in document collection. Also, the incorporation of relevance feedback in DBH based multimedia document indexing and retrieval is an interesting direction for exploration.
- Application of different performance measures such as mean reciprocal rank and discounted cumulative gain for the MKL formulation for retrieval problem. The effect of different regularization methods in MKL, and the enforcement of sparsity constraint in optical kernel learning would require new optimization framework.
- Exploration of ranking based fusion in approximate nearest neighbour based retrieval using single and multiple feature based representation. Document retrieval using multiple modalities by ranking based fusion in approximate nearest neighbour based retrieval. This would compare and establish the effectiveness of presented multi-modal retrieval methods with fusion based methods.
- Application of deeper topic learning methods such as supervised, and sequential LDA for exploring the document space, and learning the cross-modal correlation for multi-modal retrieval. Multi-modal concept based recognition by modelling the fusion of semantic concepts extracted by advance graphical models.
- Theoretical analysis of used classifiers, feature extraction mechanism for justifying the applicability for different applications. Such study would provide deeper understanding of different learning methods for a particular application scenario would

help to improve the overall performance.

Bibliography

[1] Sadegh Abbasi, Farzin Mokhtarian, and Josef Kittler, *Curvature scale space image in shape similarity retrieval*, Multimedia Systems **7** (1999), 467–476.

[2] Mausumi Acharyya and Malay K. Kundu, *Document image segmentation using wavelet scaleâŁ“space features*, IEEE Transactions on Circuits and Systems for Video Technology (2002), 1117 – 1127.

[3] Tomasz Adamek, Noel E O'Connor, and Alan F Smeaton, *Word matching using single closed contours for indexing handwritten historical documents*, International Journal on Document Analysis and Recognition **9** (2007), no. 2, 153 – 165.

[4] W. H. Adams, G. Iyengar, M. R. Naphade, C. Neti, H. J. Nock, and J. R. Smith, *Semantic indexing of multimedia content using visual, audio and text cues*, EURASIP Journal on Applied Signal Processing **2** (2003), 170–185.

[5] Charu C. Aggarwal and Philip S. Yu, *On effective conceptual indexing and similarity search in text data*, Proceeding of International Conference on Data Mining, 2001, pp. 3–10.

[6] Naif Alajlan, Ibrahim El Rube, Mohamed S. Kamel, and George Freeman, *Shape retrieval using triangle-area representation and dynamic space warping*, Pattern Recognition **40** (2007), 1911–1920.

[7] Mohand Said Allili and Djemel Ziou, *Fusion of global and local features for face verification*, Proceedings of the $!6^{th}$ International Conference on Pattern Recognition **2** (2002), 382–385.

[8] Alexandr Andoni and Piotr Indyk, *Near optimal hashing algorithms for approximate nearest neighbor in high dimensions*, Communications of the ACM **51** (2008), no. 1, 117–122.

[9] Sameer Antani and lalitha Agnihotri, *Gujarati character recognition*, Proceedings of the 5^{th} International Conference on Document Analysis and Recognition, 1999, pp. 418–421.

[10] A. Antonacopoulos, C. Clausner, C. Papadopoulos, and S. Pletschacher, *Historical document layout analysis competition*, Proceedings of the 2011 International Conference on Document Analysis and Recognition, 2011, pp. 1516–1520.

[11] Apostolos Antonacopoulos, Stefan Pletschacher, David Bridson, and Christos Papadopoulos, *Icdar 2009 page segmentation competition*, Proceedings of the 2009 10th International Conference on Document Analysis and Recognition, ICDAR '09, 2009, pp. 1370–1374.

[12] Deepak Arya, C. V. Jawahar, , Chakravorty Bhagvati, Tushar Patnaik, B. B. Chaudhuri, G. S. Lehal, Santanu Chaudhury, and A. G. Ramakrishna, *Experiences of integration and performance testing of multilingual ocr for printed indian scripts*, Proceedings of 2011 Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data, 2011, pp. 1–9.

[13] Akshay Asthana, Roland Goecke, Novi Quadrianto, and Tom Gedeon, *Learning based automatic face annotation for arbitrary poses and expressions from frontal images only*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1635–1642.

[14] Available at `http://groups.csail.mit.edu/vision/TinyImages/`.

[15] Available at `http://ocr.cdacnoida.in/`.

[16] Available at `https://www.irislink.com/`.

[17] Available at `http://www.adobe.com/`.

[18] Available at `http://www.cs.toronto.edu/kriz/cifar`.

[19] Available at `http://yann.lecun.com/exdb/mnist/`.

[20] Shuyong Bai, Linlin Li, and Chew Lim Tan, *Keyword spotting in document images through word shape coding*, Proceedings of the 10^{th} International Conference on Document Analysis and Recognition, 2009, pp. 331–335.

[21] Reena Bajaj and Santanu Chaudhury, *Signature verification using multiple neural classifiers*, Pattern Recognition **30** (1997), no. 1, 1–7.

[22] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando de Freitas, David M. Blei, and Michael I. Jordan, *Matching words and pictures*, Journal of Machine Learning Research **3** (2003), 1107–1135.

[23] Jayanta Basak, Koustav Bhattacharya, and Santanu Chaudhury, *Multiple exemplar-based facial image retrieval using independent component analysis*, IEEE Transactions on Image Processing **15** (2006), no. 12, 3773–3783.

[24] J.S. Beis and D.G. Lowe, *Shape indexing using approximate nearest-neighbour search in high-dimensional spaces*, Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on, 1997, pp. 1000–1006.

[25] Serge Belongie, Jitendra Malik, and Jan Puzicha, *Shape matching and object recognition using shape contexts*, IEEE Transaction on Pattern Analysis and Machine Intelligence **24** (2002), no. 24, 509–522.

[26] Margherita Berardi, Oronzo Altamura, Michelangelo Ceci, and Donato Malerba, *A color-based layout analysis to process censorship cards of film archives*, Proceedings of the 8^{th} International Conference on Document Analysis and Recognition, 2005, pp. 1110–1114.

[27] Anurag Bhardwaj, Srirangaraj Setlur, and Venu Govindaraju, *Sanskrit computational linguistics*, Springer-Verlag, 2009, pp. 403–416.

[28] Christian Bhm, Stefan Berchtold, and Daniel A. Keim, *Searching in high-dimensional spaces: Index structures for improving the performance of multimedia databases*, ACM Computing Surveys **33** (2001), no. 3.

[29] Manuele Bicego, Vittorio Murino, and MÃ¡rio A.T. Figueiredo, *Similarity-based classification of sequences using hidden markov models*, Pattern Recognition **37** (2004), no. 12, 2281–2291.

[30] David M. Blei and Michael I. Jordan, *Modeling annotated data*, Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, SIGIR '03, 2003, pp. 127–134.

[31] David M Blei, Andrew Y Ng, and Michael I Jordan, *Latent dirichlet allocation*, Journal of Machine Learning Research **3** (2003), 993–1022.

[32] Miroslaw Bober, *Mpeg-7 visual shape descriptors*, IEEE Transactions on Circuits and Systems and Video Technology **11** (2001), no. 6, 716–719.

[33] Christian Böhm, Stefan Berchtold, and Daniel A. Keim, *Searching in high-dimensional spaces: Index structures for improving the performance of multimedia databases*, ACM Comput. Surv. **33** (2001), no. 3, 322–373.

[34] B.S.Manjunathi and W.Y. Ma, *Texture features for browsing and retrieval of image data*, IEEE Transactions on Pattern Analysis and Machine Intelligence **18** (1996), no. 8, 837–842.

[35] H. Bunke, S. Bengio, and A. Vinciarelli, *Offline recognition of unconstrained handwritten texts using hmms and statistical language models*, Pattern Analysis and Machine Intelligence, IEEE Transactions on **26** (2004), no. 6, 709–720.

[36] Horst Bunke and Kaspar Riesen, *Recent advances in graph-based pattern recognition with applications in document analysis*, Pattern Recognition **44** (2011), no. 5, 1057 – 1067.

[37] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender, *Learning to rank using gradient descent*, Proceedings of the 22nd international conference on Machine learning, ICML '05, 2005, pp. 89–96.

[38] Andrew Busch, Wageeh W. Boles, and Sridha Sridharan, *Texture for script identification*, IEEE Transactions on Pattern Analysis and Machine Intelligence **27** (2005), 1720–1732.

[39] Yang Cao, Shuhua Wang, and Heng Li, *Skew detection and correction in document images based on straight-line fitting*, Pattern Recognition Letters **24** (2003), no. 12, 1871 – 1879.

[40] Chad Carson, Megan Thomas, Serge Belongie, Joseph M. Hellerstein, and Jitendra Malik, *Blobworld: A system for region-based image indexing and retrieval*, In Third International Conference on Visual Information Systems, Springer, 1999, pp. 509–516.

[41] R. Cattoni, S. Messelodi T. Coianiz, and C. M. Modena, *Geometric layout analysis techniques for document image understanding: a review*, Technical report, IRST (1998).

[42] S. Chanda and U. Pal, *English, devnagari and urdu text identification*, In Proceedings of International Conference on Cognition and Recognition, December 2005, pp. 538 – 545.

[43] S. Chanda, R.K. Roy, and U. Pal, *English and tamil text identification*, In Proceeding of National Conference on Recent Trends in Information Systems, July 2006, pp. 184 – 187.

[44] Moses S Charikar, *Similarity estimation techniques from rounding algorithms*, Proceedings of the 34^{th} ACM Symposium on Theory of Computing, 2002, pp. 380–388.

[45] Vasileios T. Chasanis, Aristidis C. Likas, and Nikolaos P. Galatsanos, *Scene detection in videos using shot clustering and sequence alignment*, Transactions on Multimedia **11** (2009), no. 1, 89–100.

[46] B. B. Chaudhuri and U. Pal, *A complete printed bangla ocr system*, Pattern Recognition **31** (1998), no. 5, 531 – 549.

[47] Santanu Chaudhury, Gaurav Harit, Shekar Madnani, and R.B.Shet, *Identification of scripts of indian languages by combining trainable classifiers*, Proc. of ICVGIP (2000).

[48] Santanu Chaudhury, Megha Jindal, and Sumantra Dutta Roy, *Model-guided segmentation and layout labelling of document images using a hierarchical conditional random field*, Proceedings of the 3^{rd} International Conference on Pattern Recognition and Machine Intelligence, 2009, pp. 375–380.

[49] Santanu Chaudhury, Geetika Sethi, Anand Vyas, and Gaurav Harit, *Devising interactive access techniques for indian language document images*, Proceedings of the 7^{th} International Conference on Document Analysis and Recognition, vol. 2, 2003, pp. 885–889.

[50] Santanu Chaudhury and Rabindra Sheth, *Trainable script identification strategies for indian languages*, Proceedings of the Fifth International Conference on Document Analysis and Recognition, ICDAR '99, 1999, pp. 657–.

[51] Gal Chechik, Eugene Ie, Martin Rehn, Samy Bengio, and Dick Lyon, *Large-scale content-based audio retrieval from text queries*, Proceedings of the 1st ACM international conference on Multimedia information retrieval, 2008, pp. 105–112.

[52] F. R. Chen, L. D. Wilcox, and D.S. Bloomberg, *Word spotting in scanned images using hidden markov models*, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing **5** (1993), 1–4.

[53] Atul Chhabra, *Graphic symbol recognition: An overview*, Graphics Recognition Algorithms and Systems (Karl Tombre and Atul Chhabra, eds.), Springer Berlin / Heidelberg, 1998, pp. 68–79.

[54] Alex Yong-Sang Chia, Deepu Rajan, Maylor Karhang Leung, and Susanto Rahardja, *Object recognition by discriminative combinations of line segments, ellipses, and appearance features*, IEEE Trans. Pattern Anal. Mach. Intell. **34** (2012), no. 9, 1758–1772.

[55] Young Deok Chun, Nam Chul Kim, and Ick Hoon Jang, *Content-based image retrieval using multi-resolution colour and texture features*, IEEE Transactions on Multimedia **10** (2008), no. 6, 1073–1084.

[56] Young Deok Chun, Sang Yong Seo, and Nam Chul Kim, *Image retrieval using bdip and bvlc moments*, IEEE Transactions on Circuits and Systems for Video Technology **13** (2003), no. 9, 951–957.

[57] Antonio Clavelli and Dimosthenis Karatzas, *Text segmentation in colour posters from the spanish civil war era*, Proceedings of the 10^{th} International Conference on Document Analysis and Recognition, 2009, pp. 181 – 185.

[58] Adam Coates, Blake Carpenter, Carl Case Sanjeev Satheesh, Bipin Suresh, Tao Wang, David J Wu, and Andrew Y Ng., *Text detection and character recognition in scene images with unsupervised feature learning*, Proceedings of ICDAR (2011), 440–445.

[59] C. Cotsaces, N. Nikolaidis, and I. Pitas, *Video shot detection and condensed representation. a review*, Signal Processing Magazine, IEEE **23** (2006), no. 2, 28–37.

[60] Available at `http://vision.fe.uni-lj.si/cvbase06/downloads.html` CVBASE '06 Dataset.

[61] Andrew Czarn, Cara Macnish, Kaipillil Vijayan, and Berwin Turlach, *Statistical exploratory analysis of genetic algorithms*, IEEE Transactions on Evolutionary Computation **8** (2004), 2004.

[62] Navneet Dalal and Bill Triggs, *Histograms of oriented gradients for human detection*, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1, CVPR '05, 2005, pp. 886–893.

[63] Venu Dasigi, Reinhold C. Mann, and Vladimir A. Protopopescu, *Information fusion for text classification - an experimental comparison*, Pattern Recognition **34** (2001), 2413–2425.

[64] Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S. Mirrokni, *Locality-sensitive hashing scheme based on p-stable distributions*, Proceedings of the Annual Symposium on Computational Geometry (2004), 253–662.

[65] Teofilo E. de Campos, Bodla Rakesh Babu, and Manik Varma, *Character recognition in natural images*, Proceedings of International Conference on Computer Vision Theory and Applications (VISAPP), February 2009.

[66] Kalyanmoy Deb and Samir Agrawal, *Understanding interactions among genetic algorithm parameters*, Foundations of Genetic Algorithms 5, Morgan Kaufmann, 1998, pp. 265–286.

[67] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman, *Indexing by latent semantic indexing*, Journal of the Society for Information Science **41** (1990), no. 6, 391–407.

[68] Rees. D.G., *Essential Statistics*, 4 ed., Chapman and Hall/CRC, 2001.

[69] B. V. Dhandra, P. Nagabhushan, Mallikarjun Hangarge, Ravindra Hegadi, and V. S. Malemath, *Script identification based on morphological reconstruction in document images*, International Conference on Pattern Recognition, 2006, pp. 950–953.

[70] B.V. Dhandra, H. Mallikarjun, R. Hegadi, and V.S. Malemath, *Word-wise script identification from bilingual documents based on morphological reconstruction*, Digital Information Management, 2006 1st International Conference on, December 2007, pp. 389 –394.

[71] D. Dhanya and A. G. Ramakrishnan, *Script identification in printed bilingual documents*, Proceedings of the 5^{th} International Workshop on Document Analysis Systems V, Springer-Verlag, 2002, pp. 13–24.

[72] Jignesh Dholakia, Atul Negi, and S. Rama Mohan, *Zone identification in the printed gujarati text*, Proceedings of the 8^{th} International Conference on Document Analysis and Recognition, 2005, pp. 272 – 276.

[73] Jignesh Dholakia, Archit Yajnik, and Atul Negi, *Wavelet feature based confusion character sets for gujarati script*, Proceedings of International Conference on Computational Intelligence and Multimedia Applications, vol. 2, 2007, pp. 366–370.

[74] C. S. Dima, N. Vandapel, and M. Hebert, *Classifier fusion for outdoor obstacle detection*.

[75] Nevenka Dimitrova, Hong-Jiang Zhang, Behzad Shahraray, Ibrahim Sezan, Thomas Huang, and Avideh Zakhor, *Applications of video-content analysis and retrieval*, IEEE MultiMedia **9** (2002), no. 3, 42–55.

[76] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, *Behaviour recognition via sparse spatio-temporal features*, Proceedings of the IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005, pp. 65 – 72.

[77] Ramprasath Dorairaj and Kamesh R Namuduri, *Compact combination of mpeg-7 color and texture descriptors for image retrieval*, Proceedings of the 38^{th} Asilomar Conference on Signals, Systems and Computers, vol. 1, 2004, pp. 387–391.

[78] Ritendra Dutta, Dhiraj Joshi, Jia Li, and James Z Wang, *Image retrieval: Ideas, influences, and trends of the new age*, ACM Computing Surveys **40** (2008), no. 2, 262–282.

[79] P. Duygulu, Kobus Barnard, J. F. G. de Freitas, and David A. Forsyth, *Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary*, Proceedings of the 7^{th} European Conference on Computer Vision-Part IV, Springer-Verlag, 2002, pp. 97–112.

[80] Abdulmotaleb El, Saddik Mohan, S. Kankanhalli, P. K. Atrey, M. A. Hossain, A. El Saddik, A. El Saddik, and M. S. Kankanhalli, *Multimodal fusion for multimedia analysis: a survey*, 2010, pp. 345–379.

[81] Thomas Deselaers *et al.*, *Overview of the imageclef 2007 object retrieval task*, In Working Notes of the 2007 CLEF Workshop, 2007.

[82] Boris Epshtein, Eyal Ofek, and Yonatan Wexler, *Detecting text in natural scenes with stroke width transform.*, Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2010, pp. 2963–2970.

[83] Hugo Jair Escalante, Carlos A. Hérnadez, Luis Enrique Sucar, and Manuel Montes, *Late fusion of heterogeneous methods for multimedia image retrieval*, Proceedings

of the 1st ACM international conference on Multimedia information retrieval, MIR '08, 2008, pp. 172–179.

[84] Sergio Escalera, Alicia Fornés, Oriol Pujol, Petia Radeva, Josep Lladós, and Petia Radeva, *Circular blurred shape model for multiclass symbol recognition*, IEEE Transactions on Systems, Man, and Cybernatics - Part B: Cybernatics **41** (2011), no. 2, 497–506.

[85] Sergio Escalera, Alicia Fornés, Oriol Pujol, Petia Radeva, Gemma Sánchez, and Josep Lladós, *Blurred shape model for binary and grey-level symbol recognition*, Pattern Recognition Letters **30** (2009), no. 15, 1424–1433.

[86] Martin Ester, Hans-Peter Kriegel, Jorg Sander, and Xiaowei Xu, *A density-based algorithm for discovering clusters in large spatial databases with noise*, Proceedings of the 2^{nd} International Conference on Knowledge Discovery and Data Mining, 1996, pp. 226–231.

[87] N. Ezaki, M. Bulacu, and L. Schomaker, *Text detection from natural scene images: towards a system for visually impaired persons*, Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, vol. 2, 2004, pp. 683–686.

[88] 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 **3** (1994), no. 3-4, 231–262.

[89] Christos Faloutsos and King-Ip Lin, *Fastmap: a fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets*, Proceedings of the 1995 ACM SIGMOD international conference on Management of data, SIGMOD '95, ACM, 1995, pp. 163–174.

[90] Jianping Fan, Yuli Gao, and Hangzai Luo, *Hierarchical classification for automatic image annotation*, Proceedings of the 30^{th} annual international ACM SIGIR conference on Research and development in information retrieval, 2007, pp. 111–118.

[91] Kuo-Chin Fan, Yuan-Kai Wang, and Tsann-Ran Lay, *Marginal noise removal of document images*, Pattern Recognition **35** (2002), no. 11, 2593 – 2611.

[92] Fabio F. Faria, Adriano Veloso, Humberto M. Almeida, Eduardo Valle, Ricardo da S. Torres, Marcos A. Gonçalves, and Wagner Meira, Jr., *Learning to rank for content-based image retrieval*, Proceedings of the international conference on Multimedia information retrieval, MIR '10, 2010, pp. 285–294.

[93] Faisal Farooq, Anurag Bhardwaj, and Venu Govindaraju, *Using topic models for ocr correction*, Interneational Journal of Document Analysis and Recognition **12** (2009), 153–164.

[94] N. Fatemi, M. Lalmas, and T. Rolleke, *How to retrieve multimedia documents described by mpeg-7*, Advanced Information Systems Engineering, 16th International Conference, CAiSE04 (2004).

[95] J. Favata, G. Srikantan, and S. Srihari, *Hand printed character/digit recognition using a multiple feature/resolution phitosophy*, Proceedings of the International Workshop on on Frontiers in Handwriting Recognition, 1994, pp. 57–66.

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

[97] Rob Fergus, Yair Weiss, and Antonio Torralba, *Semi-supervised learning in gigantic image collections*, Proceedings of the Advances in Neural Information Processing Systems (2009), 522–530.

[98] Alfio Ferrara, Luca A. Ludovico, Stefano Montanelli, Silvana Castano, and Goffredo Haus, *A semantic web ontology for context-based classification and retrieval of music resources*, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCAP) **2** (2006), no. 3, 177–198.

[99] Andreas Fischer, Andreas Keller, Volkmar Frinken, and Horst Bunke, *Lexicon-free handwritten word spotting using character {HMMs}*, Pattern Recognition Letters **33** (2012), no. 7, 934 – 942.

[100] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Qian 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 **28** (1995), no. 9, 23–32.

[101] D. Frank Hsu and Isak Taksa, *Comparing rank and score combination methods for data fusion in information retrieval*, Inf. Retr. **8** (2005), no. 3, 449–480.

[102] Yoav Freund, Raj Iyer, Robert E. Schapire, and Yoram Singer, *An efficient boosting algorithm for combining preferences*, Journal of Machine Learning Research **4** (2003), 933–969.

[103] V. Frinken, A. Fischer, H. Bunke, and R. Manmatha, *Adapting blstm neural network based keyword spotting trained on modern data to historical documents*, Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on, 2010, pp. 352–357.

[104] V. Frinken, A. Fischer, R. Manmatha, and H. Bunke, *A novel word spotting method based on recurrent neural networks*, Pattern Analysis and Machine Intelligence, IEEE Transactions on **34** (2012), no. 2, 211–224.

[105] Andrea Frome, Yoram Singer, Fei Sha, and Jitendra Malik, *Learning globally-consistent local distance functions for shape-based image retrieval and classification*, IEEE 11th International Conference on Computer Vision, 2007, pp. 1–8.

[106] George Gagaudakis and Paul L. Rosin, *Incorporating shape into histograms for cbir*, Pattern Recognition **35** (2002), 81–91.

[107] J. Gao and Lei Yang, *An adaptive algorithm for text detection from natural scenes*, Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, vol. 2, 2001, pp. 84–89.

[108] U. Garain and B. B. Chaudhuri, *Recognition of online handwritten mathematical expressions*, IEEE Transactions on System, Man and Cybernetics: Part B **34** (2004), no. 6, 2366–2376.

[109] U. Garain and B.B. Chaudhuri, *Segmentation of touching characters in printed devnagari and bangla scripts using fuzzy multifactorial analysis*, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews **32** (2002), no. 4, 449 –459.

[110] Basilis Gatos and Ioannis Pratikakis, *Segmentation-free word spotting in historical printed documents*, Proceedings of the 2009 10th International Conference on Document Analysis and Recognition, ICDAR '09, 2009, pp. 271–275.

[111] Peter Gehler and Sebastian Nowozin, *On feature combination for multiclass object classification*, Proceedings of the IEEE International Conference on Computer Vision, 2009, pp. 1–8.

[112] Michel X. Goemans and David P. Williamson, *Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming*, Journal of the ACM **42** (1995), no. 6, 1115–1145.

[113] David E. Goldberg, *Genetic algorithms in search, optimization and machine learning*, Addison Wesley Longman (Singapore) Private Limited, 2000.

[114] Mehmet Gonen and Ethem Alpaydm, *Multiple kernel learning algorithms*, Journal of Machine Learning Research, vol. 12, 2011, pp. 2211–2268.

[115] Yihong Gong, G. Proietti, and C. Faloutsos, *Image indexing and retrieval based on human perceptual color clustering*, Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on, 1998, pp. 578–583.

[116] Google Inc., *Book search dataset*, August 2007, Version 1.

[117] Philippe Henri Gosselin, Matthieu Cord, and Sylvie Philipp-Foliguet, *Combining visual dictionary, kernel-based similarity and learning strategy for image category retrieval*, Computer Vision and Image Understanding **110** (2008), 403âŁ“–417.

[118] Hideaki Goto, *Redefining the dct-based feature for scene text detection*, International Journal of Document Analysis and Recognition (IJDAR) **11** (2008), no. 1, 1–8.

[119] Venu Govindaraju and Srirangaraj Setlur (eds.), *Guide to ocr for indic scripts*, Document Recognition and Retrieval Series: Advances in Computer Vision and Pattern Recognition, Springer-Verlag London Limited, 2009.

[120] David Grangier and Samy Bengio, *A discriminative kernel-based approach to rank images from text queries*, IEEE Transactions on Pattern Analysis and Machine Intelligence **30** (2008), no. 8, 1371–1384.

[121] Cosmin Grigorescu and Nicolai Petkov, *Distance sets for shape filters and shape recognition*, IEEE Transactions on Image Processing **12** (2003), no. 10, 1274–1286.

[122] Simona E. Grigorescu, Nicolai Petkov, and Peter Kruizinga, *Comparison of texture features based on gabor filters*, IEEE Transactions on Image Processing (2002), 1160 – 1167.

[123] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, *Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation*, Computer Vision, 2009 IEEE 12th International Conference on, 2009, pp. 309–316.

[124] D. Guo and S. Z. Li, *Content-based audio classification and retrieval by support vector machines*, IEEE Transactions on Neural Networks **14** (2003), no. 1, 209–215.

[125] Maya R. Gupta, Nathaniel P. Jacobson, and Eric K. Garcia, *Ocr binarization and image pre-processing for searching historical documents*, Pattern Recognition **40** (2007), no. 2, 389–397.

[126] Parisa Haghani, Sebastian Michel, and Karl Aberer, *Distributed similarity search in high dimensions using locality sensitive hashing*, Proceedings of the 12^{th} International Conference on Extending Database Technology, 2009, pp. 744–755.

[127] Jingrui He, Mingjing Li, Hong-Jiang Zhang, Hanghang Tong, and Changshui Zhang, *Manifold-ranking based image retrieval*, Proceedings of the 12th annual ACM international conference on Multimedia, MULTIMEDIA '04, 2004, pp. 9–16.

[128] Gregor Heinrich, *Parameter estimation for text analysis*, Tech. report, 2004.

[129] Avishai Hendel, Daphna Weinshall, and Shmuel Peleg, *Identifying surprising events in videos using bayesian topic models*, Proceedings of the 10^{th} Asian conference on Computer vision, vol. Part III, 2011, pp. 448–459.

[130] Jae-Pil Heo, Youngwoon Lee, Junfeng He, Shih-Fu Chang, and Sung-Eui Yoon, *Spherical hashing*, Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, 2012, pp. 2957–2964.

[131] Ralf Herbrich, Thore Graepel, and Klaus Obermayer, *Support vector learning for ordinal regression*, Proceedings of the International Conference on Artificial Neural Networks, 1999, pp. 97–102.

[132] Tomer Hertz, Aharon Bar-Hillel, and Daphna Weinshall, *Learning distance functions for image retrieval*, Proceedings of the 2004 IEEE computer society conference on Computer vision and pattern recognition, CVPR'04, 2004, pp. 570–577.

[133] Hery Heryanto, Saiful Akbar, and Benhard Sitohang, *Direct access in content-based audio information retrieval: A state of the art and challenges.*, 2011, pp. 1–6.

[134] J. Hochberg, P. Kelly, T. Thomas, and L. Kerns, *Automatic script identification from document images using cluster-based templates*, IEEE Transactions on Pattern Analysis and Machine Intelligence **19** (1997), no. 2, 176 –181.

[135] Thomas Hofmann, *Probabilistic latent semantic indexing*, Proceedings of the 22^{nd} annual international ACM SIGIR conference on Research and development in information retrieval, 1999, pp. 50–57.

[136] Steven C.h. Hoi, Wei Liu, and Shih-Fu Chang, *Semi-supervised distance metric learning for collaborative image retrieval and clustering*, ACM Trans. Multimedia Comput. Commun. Appl. **6** (2010), no. 3, 18:1–18:26.

[137] Timothy Hospedales, Shaogang Gong, and Tao Xiang, *A markov clustering topic model for mining behaviour in video*, Proceedings of the International Conference on Computer Vision, 2009, pp. 1165 –1172.

[138] Nicholas R. Howe, Shaolei Feng, and R. Manmatha, *Finding words in alphabet soup: Inference on freeform character recognition for historical scripts*, Pattern Recognition **42** (2009), no. 12, 3338–3347.

[139] Nicholas R. Howe, Toni M. Rath, and R. Manmatha, *Boosted decision trees for word recognition in handwritten document retrieval*, Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '05, 2005, pp. 377–383.

[140] E. Indermuhle, V. Frinken, and H. Bunke, *Mode detection in online handwritten documents using blstm neural networks*, Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on, 2012, pp. 302–307.

[141] Piotr Indyk and Rajeev Motwani, *Approximate nearest neighbor - towards removing the curse of dimensionality*, Proceedings of the 30^{th} ACM Symposium on Theory of Computing (1998), 604–613.

[142] Luarent Itti and Christof Koch, *Feature combination strategies for saliency-based visual attention systems*, Journal of Electronic Imaging **10** (2001), no. 1, 161–169.

[143] Nisha Jain, Santanu Chaudhury, Sumatra Dutta Roy, P. Mukerjee, K. Seal, and K. Talluri, *A novel learning-based framework for detecting interesting events in soccer videos*, Proceedings of the 6^{th} Indian Conference on Computer Vision, Graphics Image Processing, dec. 2008, pp. 119 –125.

[144] Vidit Jain and Manik Varma, *Learning to re-rank: query-dependent image re-ranking using click data*, Proceedings of the 20th international conference on World wide web, WWW '11, 2011, pp. 277–286.

[145] C. V. Jawahar, , M. N. S. S. K. Pavan Kumar, and S. S. Ravi Kiran, *A bilingual ocr for hindi-telugu documents and its applications*, Proceedings of the Seventh International Conference on Document Analysis and Recognition - Volume 1, ICDAR '03, IEEE Computer Society, 2003.

[146] Herve Jegou, Matthijs Douze, and Cordelia Schmid, *Hamming embedding and weak geometric consistency for large scale image search*, Proceedings of the 10th European Conference on Computer Vision: Part I, ECCV '08, 2008, pp. 304–317.

[147] J. Jeon, V. Lavrenko, and R. Manmatha, *Automatic image annotation and retrieval using cross-media relevance models*, Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, 2003, pp. 119–126.

[148] Yangqing Jia, Mathieu Salzmann, and Trevor Darrell, *Learning cross modality similarity for multinomial data*, Proceedings of the IEEE International Conference on Computer Vision (2011), 2407–2414.

[149] Yu-Gang Jiang and Chong-Wah Ngo, *Visual word proximity and linguistics for semantic video indexing and near-duplicate retrieval*, Computer Vision and Image Understanding **113** (2009), no. 3, 405–414.

[150] Feng Jing, Mingjing Li, Hong-Jiang Zhang, and Bo Zhang, *An efficient and effective region-based image retrieval framework*, Image Processing, IEEE Transactions on **13** (2004), no. 5, 699–709.

[151] A. L. Kesidis and B. Gatos, *Efficient cut-off threshold estimation for word spotting applications*, Document Analysis and Recognition (ICDAR), 2011 International Conference on, 2011, pp. 279–283.

[152] Wolf Kienzle, Bernhard Schölkopf, Felix A. Wichmann, and Matthias O. Franz, *How to find interesting locations in video: a spatiotemporal interest point detector learned from human eye movements*, Proceedings of the 29^{th} DAGM conference on Pattern recognition, 2007, pp. 405–414.

[153] Gyeonghwan Kim and V. Govindaraju, *A lexicon driven approach to handwritten word recognition for real-time applications*, Pattern Analysis and Machine Intelligence, IEEE Transactions on **19** (1997), no. 4, 366–379.

[154] Jongryeol Kim, Kukhwan Seo, and Kyusik Chung, *A systematic approach to classifier selection on combining multiple classifiers for handwritten digit recognition*, Proceedings of the 4th International Conference on Document Analysis and Recognition, 1997, pp. 459–462.

[155] Kwang In Kim, Keechul Jung, and Jin-Hyung Kim, *Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm*, Pattern Analysis and Machine Intelligence, IEEE Transactions on **25** (2003), no. 12, 1631–1639.

[156] T. Konidaris, B. Gatos, K. Ntzios, I. Pratikakis, S. Theodoridis, and S.J. Perantonis, *Keyword-guided word spotting in historical printed documents using synthetic data and user feedback*, International Journal of Document Analysis and Recognition (IJDAR) **9** (2007), no. 2-4, 167–177.

[157] Praveen Krishnan, Ravi Shekhar, and C. V. Jawahar, *Content level access to digital library of india pages*, Proceedings of the Eighth Indian Conference on Computer Vision, Graphics and Image Processing, 2012, pp. 5:1–5:8.

[158] Brian Kulis and Kristen Grauman, *Kernelized locality sensitive hashing for scalable image search*, Proceedings of the IEEE International Conference on Computer Vision, 2009, pp. 2130–2137.

[159] A. Kumar and D. Zhang, *Personal recognition using hand shape and texture*, Trans. Img. Proc. **15** (2006), no. 8, 2454–2461.

[160] Anand Kumar, C V Jawahar, and R Manmatha, *Efficient search in document image collections*, Proceedings of the 8^{th} Asian Conference on Computer Vision, 2007, pp. 586–595.

[161] Sunil Kumar, Rajat Gupta, Nitin Khanna, Santanu Chaudhury, and Shiv Dutt Joshi, *Text extraction and document image segmentation using matched wavelets and mrf model*, IEEE Transactions on Image Processing **8** (2007), 2117–2128.

[162] Ludmila I. Kuncheva, *A theoretical study on six classifier fusion strategies*, IEEE Transactions on Pattern Analysis and Machine Intelligence **24** (2002), no. 2, 281–286.

[163] Caroline Lacoste, Joo-Hwee Lim, Jean-Pierre Chevallet, and Diem Thi Hoang Le, *Medical-image retrieval based on knowledge assisted text and image indexing*, IEEE Transactions on Circuits and Systems for Video Technology **17** (2007), no. 7, 889–900.

[164] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira, *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*, Proceedings of the 8^{th} International Conference on Machine Learning, ICML '01, 2001, pp. 282–289.

[165] Stephen W. Lam, *A local-to-global approach to complex document layout analysis*, IAPR Workshop on Machine Vision Applications (1994), 13–15.

[166] Zhen-zhong Lan, Lei Bao, Shoou-I Yu, Wei Liu, and Alexander G. Hauptmann, *Double fusion for multimedia event detection*, Proceedings of the 18th international conference on Advances in Multimedia Modeling, 2012, pp. 173–185.

[167] Gert R. G. Lanckriet, Tijl De Bie, Nello Cristianini, Michael I. Jordan, and William Stafford Noble, *A statistical framework for genomic data fusion*, Bioinformatics **20** (2004), no. 16, 2626–2635.

[168] Gert R.G. Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I. Jordan, *Learning the kernel matrix with semi-definite programming*, Journal of Machine Learning Research **4** (2004), 27–72.

[169] Ivan Laptev, *On space-time interest points*, International Journal of Computer Vision **64** (2005), 107–123.

[170] Ivan Laptev, Marcin MarszaÅ,ek, Cordelia Schmid, and Benjamin Rozenfeld, *Learning realistic human actions from movies*, Proceedings of the International Conference on Computer Vision and Pattern Recognition, 2008, pp. 1 –8.

[171] Kevin Laven, Scott Leishman, and Sam Roweis, *A statistical learning approach to document image analysis*, In 8th International Conference on Document Analysis and Recognition, 2005, pp. 357–361.

[172] Svetlana Lazebnik, Cordelia Schmid, , and Jean Ponce, *A sparse texture representation using local affine regions*, IEEE Transactions on Pattern Analysis and Machine Intelligence **27** (2005), no. 8, 1265–1278.

[173] Yann Lecun, Leon Bottou, Yoshua Bengio, and Patrick Haffner, *Gradient based learning applied to document recognition*, In the Proceedings of IEEE **86** (1998), no. 11, 2278–2324.

[174] Jung-Jin Lee, Pyoung-Hean Lee, Seong-Whan Lee, Alan Yuille, and Christof Koch, *Adaboost for text detection in natural scene*, Proceedings of IEEE ICDAR (2011), 429–434.

[175] Michael S. Lew, Nicu Sebe, Chabane Djeraba, and Ramesh Jain, *Content-based multimedia information retrieval: State of the art and challenges*, ACM Transactions on Multimedia Computing, Communications and Applications **2** (2006), no. 1, 1–19.

[176] Hang Li, *Learning to Rank for Information Retrieval and Natural Language Processing*, Morgan & Claypool Publishers, 2011.

[177] Huiping Li, D. Doermann, and O. Kia, *Automatic text detection and tracking in digital video*, Image Processing, IEEE Transactions on **9** (2000), no. 1, 147–156.

[178] Piji Li, Lei Zhang, and Jun Ma, *Dual-ranking for web image retrieval*, Proceedings of the ACM International Conference on Image and Video Retrieval, CIVR '10, 2010, pp. 166–173.

[179] Xirong Li, C. G M Snoek, and Marcel Worring, *Learning social tag relevance by neighbor voting*, Multimedia, IEEE Transactions on **11** (2009), no. 7, 1310–1322.

[180] R. Lienhart and A. Wernicke, *Localizing and segmenting text in images and videos*, Circuits and Systems for Video Technology, IEEE Transactions on **12** (2002), no. 4, 256–268.

[181] M-W Lin, J-R Tapamo, and B Ndovie, *A texture-based method for document segmentation and classiï¬'cation*, In Joint Special Issue âŁ" Advances in end-user data-mining techniques (2006), 49 – 56.

[182] Yingqiang Lin and Bir Bhanu, *Evolutionary feature synthesis for object recognition*, IEEE Transactions on Systems, Man and Cybernetics-Part C: Applications and Reviews **35** (2005), no. 2, 156–171.

[183] ———, *Object detection via feature synthesis using mdl-based genetic programming*, IEEE Transactions on Systems, Man and Cybernetics-Part B: Cybernetics **35** (2005), no. 3, 538–547.

[184] Oskar Linde and Tony Lindeberg, *Object recognition using composed receptive field histograms of higher dimensionality*, Proceedings of the 17^{th} International Conference Pattern Recognition, 2004, pp. 1–6.

[185] Tie-Yan Liu, *Learning to rank for information retrieval*, Found. Trends Inf. Retr. **3** (2009), no. 3, 225–331.

[186] Wei Liu, Jun Wang, Rongrong Ji, Yu-Gang Jiang, and Shih-Fu Chang, *Supervised hashing with kernels*, Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, 2012, pp. 2074–2081.

[187] Xiaoqing Liu and J. Samarabandu, *Multiscale edge-based text extraction from complex images*, Multimedia and Expo, 2006 IEEE International Conference on, 2006, pp. 1721–1724.

[188] Ying Liu, Dengsheng Zhang, and Guojun Lu, *Region-based image retrieval with high-level semantics using decision tree learning*, Pattern Recognition **41** (2008), no. 8, 2554–2570.

[189] Ying Liu, Dengsheng Zhanga, Guojun Lua, and Wei Ying Ma, *A survey of content-based image retrieval with high-level semantics*, Pattern Recognition **40** (2007), 262–282.

[190] Josep Lladós and Gemma Sánchez, *Indexing historical documents by word shape signatures*, Proceedings of the 9^{th} International Conference on Document Analysis and Recognition, 2007, pp. 362–366.

[191] Josep Lladós, Ernest Valveny, Gemma Sánchez, and Enric Martí, *Symbol recognition: Current advances and perspectives*, Selected Papers from the 4^{th} International Workshop on Graphics Recognition Algorithms and Applications, Springer-Verlag, 2002, pp. 104–127.

[192] David G. Lowe, *Object recognition from local scale-invariant features*, Proceedings of the International Conference on Computer Vision, vol. 2, 1999, pp. 1150 –1157.

[193] Shijian Lu, Linlin Li, and Chew Lim Tan, *Document image retrieval through word shape coding*, IEEE Transactions on Pattern Analysis and Machine Intelligence **30** (2008), no. 11, 1913–1918.

[194] Qin Lv, William Josephson, Zhe Wang, Moses Charikar, and Kai Li, *Multi-probe lsh: Efficient indexing for high-dimensional similarity search*, Proceedings of the 33^{th} International Conference on Very Large Data Bases (2007), 950–961.

[195] Sriganesh Madhvanath and Venu Govindaraju, *The role of holistic paradigms in handwritten word recognition*, IEEE Transactions on Pattern Analysis and Machine Intelligence **23** (2001), no. 2, 149 – 164.

[196] Angshul Majumdar, *Bangla basic character recognition using digital curvelet transform*, Journal of Pattern Recognition Research **1** (2007), 17–26.

[197] Ameesh Makadia, Vladimir Pavlovic, and Sanjiv Kumar, *A new baseline for image annotation*, Proceedings of the 10th European Conference on Computer Vision: Part III, 2008, pp. 316–329.

[198] R Manmatha, Chengfeng Han, E M Riseman, and W B Croft, *Indexing handwriting using word matching*, Proceedings of the 1^{th} ACM International Conference on Digital Libraries (1996), 151–159.

[199] R Manmatha and Toni M Rath, *Indexing of handwritten historical documents - recent progress*, Proceedings of the Symposium on Document Image Understanding (SDIUT-03) (2003), 77–85.

[200] Christopher D. Manning, Prabhakar Raghavan, and Hinrich SchÃ¼tze, *An introduction to information retrieval*, Cambridge University Press, Cambridge, 2009.

[201] Song Mao, Azriel Rosenfeld, and Tapas Kanungo, *Document structure analysis algorithms: a literature survey*, Proceedings of SPIE Electronic Imaging (2003), 197âŁ“207.

[202] Simone Marinai, *Introduction to document analysis and recognition*, Machine Learning in Document Analysis and Recognition, 2008, pp. 1–20.

[203] Simone Marinai, *Text retrieval from early printed books*, International Journal on Document Analysis and Recognition (IJDAR) **14** (2011), no. 2, 117–129.

[204] Simone Marinai, Emanuele Marino, and Giovanni Soda, *Indexing and retrieval of words in old documents*, Proceedings of the Seventh International Conference on Document Analysis and Recognition - Volume 1, ICDAR '03, 2003, pp. 223–228.

[205] Simone Marinai, Emanuele Marino, and Giovanni Soda, *Font adaptive word indexing of modern printed documents*, IEEE Transactions on Pattern Analysis and Machine Intelligence **28** (2006), no. 8, 1187–1199.

[206] Bogdan Matei, Ying Shan, Harpreet S. Sawhney, Yi Tan, Rakesh Kumar, Daniel Huber, and Martial Hebert, *Rapid object indexing using locality sensitive hashing and joint 3d-signature space estimation*, IEEE Transactions on Pattern Analysis and Machine Intelligence **28** (2006), no. 7, 1111 – 1126.

[207] Kieran McDonald and Alan F. Smeaton, *A comparison of score, rank and probability-based fusion methods for video shot retrieval*, Proceedings of the 4th international conference on Image and Video Retrieval, CIVR'05, 2005, pp. 61–70.

[208] Carlo Meghini, Fabrizio Sebastiani, and Umberto Straccia, *A model of multimedia information retrieval*, Journal of the ACM **58** (2001), no. 5, 903–970.

[209] Tanveer Syeeda Mehmod, *Indexing of handwritten document images*, Proceedings of the 1997 Workshop on Document Image Analysis, 1997, pp. 66–73.

[210] R. Mehran, A. Oyama, and M. Shah, *Abnormal crowd behavior detection using social force model*, Proceedings of the International Conference on Computer Vision and Pattern Recognition, june 2009, pp. 935 –942.

[211] M. Merler, Rong Yan, and J.R. Smith, *Imbalanced rankboost for efficiently ranking large-scale image/video collections*, Computer Vision and Pattern Recognition, IEEE International Conference on (2009), 2607–2614.

[212] P Mermelstein, *Distance measures for speech recognition, psychological and instrumental*, vol. 116, pp. 374–388, Academic, 1976.

[213] Million Meshesha and C. V. Jawahar, *Matching word images for content-based retrieval from printed document images*, International Journal of Document Analysis and Recognition **11** (2008), 29–38.

[214] Alberto Messina and Maurizio Montagnuolo, *A generalised cross-modal clustering method applied to multimedia news semantic indexing and retrieval*, Proceedings of the 18th international conference on World wide web (New York, NY, USA), ACM, 2009, pp. 321–330.

[215] D. Mighlani, A. Hennig, N. Sherkat, and R. J. Whitrow, *A visual vocabulary for flower classification*, Proceedings of IEEE TENCON '97. Speech and Image Technologies for Computing and Telecommunications., 1997, pp. 191 – 194.

[216] Yang Mingqiang, Kpalma Kidiyo, and Ronsin Joseph, *Pattern recognition techniques, technology and applications*, ch. 3, pp. 43–90, In-Teh, Croatia, 2008.

[217] Shan Mo and V. John Mathews, *Adaptive, quadratic preprocessing of document images for binarization*, IEEE Transactions on Image Processing **7** (1998), 992–999.

[218] Florent Monay and Daniel Gatica-Perez, *Plsa-based image auto-annotation: constraining the latent space*, Proceedings of the 12^{th} annual ACM international conference on Multimedia, MULTIMEDIA '04, 2004, pp. 348–351.

[219] Pedro J. Moreno, Purdy P. Ho, and Nuno Vasconcelos, *A kullback-leibler divergence based kernel for svm classification in multimedia applications*, Proceedings of the Advances in Neural Information Processing Systems (Sebastian Thrun, Lawrence K. Saul, and Bernhard SchÃ¶lkopf, eds.), MIT Press, 2003.

[220] Yadong Mu, Jialie Shen, and Shuicheng Yan, *Weakly-supervised hashing in kernel space*, Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2010, pp. 3344–3351.

[221] Marius Muja and David G. Lowe, *Fast matching of binary features*, Proceedings of the 2012 Ninth Conference on Computer and Robot Vision, CRV '12, 2012, pp. 404–410.

[222] Dipti Prasad Mukherjee and Scott T. Acton, *Document page segmentation using multiscale clustering*, Proceedings of International Conference on Image Processing, vol. 1, 1999, pp. 234–238.

[223] Takehiro Nakayama, *Content-oriented categorization of document images*, Proceedings of the 16^{th} International Conference on Computational Linguistics **2** (1996), 818–823.

[224] Apostol (Paul) Natsev, Alexander Haubold, Jelena Tešić, Lexing Xie, and Rong Yan, *Semantic concept-based query expansion and re-ranking for multimedia retrieval*, Proceedings of the 15th international conference on Multimedia, MULTIMEDIA '07, 2007, pp. 991–1000.

[225] L. Neumann and J. Matas, *Real-time scene text localization and recognition*, Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 3538–3545.

[226] Maria-Elena Nilsback and Andrew Zisserman, *A visual vocabulary for flower classification*, Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006, pp. 1447 – 1454.

[227] David Nister and Henrik Stewenius, *Scalable recognition with a vocabulary tree*, Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2, CVPR '06, 2006, pp. 2161–2168.

[228] Il-Seok Oh, Jin-Seon Lee, and Ching Y. Suen, *Analysis of class separation and combination of class-dependent features for handwriting recognition*, IEEE Transactions on Pattern Analysis and Machine Intelligence **21** (1999), no. 10, 1089–1094.

[229] Aude Oliva and Antonio Torralba, *Modeling the shape of the scene: A holistic representation of the spatial envelope*, International Journal of Computer Vision **42** (2001), no. 3, 145–175.

[230] Ximena Olivares, Massimiliano Ciaramita, and Roelof van Zwol, *Boosting image retrieval through aggregating search results based on visual annotations*, Proceedings of the 16th ACM international conference on Multimedia, MM '08, 2008, pp. 189–198.

[231] Andreas Opelt, Axel Pinz, Michael Fussenegger, and Peter Auer, *Generic object recognition with boosting*, IEEE Trans. Pattern Anal. Mach. Intell. **28** (2006), no. 3, 416–431.

[232] Nobuyuki Otsu, *A threshold selection method from gray-level histograms*, IEEE Transactions on Systems, Man, and Cybernetics **9** (1979), no. 1, 62–66.

[233] M. C. Padma and P. A. Vijya, *Script identification from trilingual documents using profile based segmentation*, International Journal of Computer Science and Applications, vol. 7, 2010, pp. 16–33.

[234] M.C. Padma and P. Nagabhushan, *Identification and separation of text words of kannada, hindi and english languages through discriminating features*, In Proceeding of National Conference on Document Analysis and Recognition, July 2003, pp. 252 – 260.

[235] U. Pal and B.B. Chaudhuri, *Ocr in bangla: an indo-bangladeshi language*, Pattern Recognition, 1994. Vol. 2 - Conference B: Computer Vision Image Processing., Proceedings of the 12th IAPR International. Conference on, vol. 2, oct 1994, pp. 269 –273.

[236] Umapada Pal and B. B. Chaudhuri, *Script line separation from indian multi-script documents*, Proceedings of the 5^{th} International Conference on Document Analysis and Recognition, 1999, pp. 406–409.

[237] Umapada Pal and B.B. Chaudhuri, *Automatic separation of words in multi-lingual multi-script indian documents*, Proceedings of Fourth International Conference on Document Analysis and Recognition **2** (1997), 576 – 579.

[238] ———, *Identification of different script lines from multi-script documents*, Image and Vision Computing **20** (2002), no. 13 - 14, 945 – 954.

[239] Umapada Pal, Partha Pratim Roy, Nilamadhaba Tripathy, and Josep Lladós, *Multi-oriented bangla and devnagari text recognition*, Pattern Recognition **43** (2010), no. 12, 4124 – 4136.

[240] Umapada Pal, Suranjit Sinha, and B. B. Chaudhuri, *Multi-script line identification from indian document*, Proceedings of the 7^{th} International Conference on Document Analysis and Recognition, 2003, pp. 880–884.

[241] Peeta Basa Pati and A. G. Ramakrishnan, *Word level multi-script identification*, Pattern Recognition Letters **29** (2008), 1218–1229.

[242] S. Basavaraj Patil and N. V. Subbareddy, *Neural network based system for script identification in indian documents*, Sadhana-academy Proceedings in Engineering Sciences **27** (2002), 83–97.

[243] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, *Object retrieval with large vocabularies and fast spatial matching*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007.

[244] M. Pijl, S. van de Par, and Caifeng Shan, *An event-based approach to multi-modal activity modelling and recognition*, Proceedings of the International Conference on Pervasive Computing and Communications, April 2010, pp. 98 –106.

[245] John C. Platt, Nello Cristianini, and John Shawe Taylor, *Large margin dags for multiclass classification*, Proceedings of the Advances in Neural Information Processing Systems, vol. 12, MIT Press, Cambridge, 2000, pp. 547–553.

[246] M.S. Praveen, K.P. Sankar, and C.V. Jawahar, *Character n-gram spotting in document images*, Document Analysis and Recognition (ICDAR), 2011 International Conference on, 2011, pp. 941–945.

[247] Xiaojun Qi and Yutao Han, *A novel fusion approach to content-based image retrieval*, Pattern Recognition (2005), no. 38, 2449–2465.

[248] ———, *Incorporating multiple svms for automatic image annotation*, Pattern Recognition **40** (2007), 728–741.

[249] Maxim Raginsky and Svetlana Lazebnik, *Locality-sensitive binary codes from shift-invariant kernels*, Advances in Neural Information Processing Systems 22 (Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, eds.), 2009, pp. 1509–1517.

[250] Rouhollah Rahmani, Sally A. Goldman, Hui Zhang, John Krettek, and Jason E. Fritts, *Localized content based image retrieval*, Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval, MIR '05, 2005, pp. 227–236.

[251] ShyamSundar Rajaram, Charlie K. Dagli, Nemanja Petrovic, and Thomas S. Huang, *Diverse active ranking for multimedia search.*, CVPR, IEEE Computer Society, 2007.

[252] Alain Rakotomamonjy, Francis bach, Stephane Canu, and Yves Grandvalet, *More efficiency in multiple kernel learning*, Proceedings of the International Conference on Machine Learning, vol. 772, 2007, pp. 775–782.

[253] Eduardo H. Ramirez and Ramon F. Brena, *An information-theoretic approach for unsupervised topic mining in large text collections*, Proceedings of the International Joint Conference on Web Intelligence and Intelligent Agent Technology, 2009, pp. 331–334.

[254] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert R.G. Lanckriet, Roger Levy, and Nuno Vasconcelos, *A new approach to cross-modal multimedia retrieval*, Proceedings of the international conference on Multimedia, MM '10, ACM, 2010, pp. 251–260.

[255] Toni M. Rath and R. Manmatha, *Features for word spotting in historical manuscripts*, Proceedings of the 7^{th} International Conference on Document Analysis and Recognition, vol. 1, 2003, pp. 218–222.

[256] Toni M Rath and R Manmatha, *Word image matching using dynamic time warping*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, June 2003, pp. 521–527.

[257] Toni M. Rath, R. Manmatha, and Victor Lavrenko, *A search engine for historical manuscript images*, Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, 2004, pp. 369–376.

[258] TonyM. Rath and R. Manmatha, *Word spotting for historical documents*, International Journal of Document Analysis and Recognition (IJDAR) **9** (2007), no. 2-4, 139–152.

[259] Mika Rautiainen, Timo Ojala, and Tapio Seppänen, *Analysing the performance of visual, concept and text features in content-based video retrieval*, Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval, MIR '04, 2004, pp. 197–204.

[260] Mika Rautiainen and T. Seppdnen, *Comparison of visual features and fusion techniques in automatic detection of concepts from news video*, 2005, pp. 932–935.

[261] Ajoy Kumar Ray and B. Chatterjee, *Design of a nearest neighbor classifier system for bengali character recognition*, Journal of IETE **30** (1984), no. 6, 226 – 229.

[262] Wei Xin Ren, Sameer Singh, Maneesha Singh, and Y S Zhu, *State-of-the-art on spatio-temporal information-based video retrieval*, Pattern Recognition **42** (2009), no. 2, 267–282.

[263] José A. Rodríguez-Serrano and Florent Perronnin, *Handwritten word-spotting using hidden markov models and universal vocabularies*, Pattern Recogn. **42** (2009), no. 9, 2106–2116.

[264] Azriel Rozenfeld and John L. Pflatz, *Sequential operations in digital picture processing*, Journal of Association for Computing Machinery **13** (1966), no. 4, 471–494.

[265] Dymitr Ruta and Bogdan Gabrys, *An Overview of Classifier Fusion Methods*, (2000).

[266] Hanan Samet, *Foundations of multidimensional and metric data structures (the morgan kaufmann series in computer graphics and geometric modeling)*, Morgan Kaufmann Publishers, San Francisco, CA, USA, 2005.

[267] Pramod K. Sankar and C. V. Jawahar, *Enabling search over large collections of telugu document images âŁ“ an automatic annotation based approach*, Computer Vision, Graphics and Image Processing (Prem Kalra and Shmuel Peleg, eds.), Lecture Notes in Computer Science, vol. 4338, Springer, 2006, pp. 837–848.

[268] J. Sauvola and M. PietikÃ¤inen, *Adaptive document image binarization*, Pattern Recognition **33** (2000), no. 2, 225 – 236.

[269] Ediz Saykol, Ali Kemal Sinop, Ugur Gudukbay, Ozgur Ulusoy, and A. Enis Cetin, *Content-based retrieval of historical ottoman documents stored as textual images*, IEEE Transactions on Image Processing **13** (2004), no. 3, 314–325.

[270] F. Scalzo, G. Bebis, M. Nicolescu, L. Loss, and A. Tavakkoli, *Feature fusion hierarchies for gender classification*, Proceedings of International Conference on Pattern Recognition, 2008, pp. 1–4.

[271] Bernhard Scholkopf and Alexander J. Smola, *Learning with kernels: Support vector machines, regularization, optimization, and beyond*, The MIT Press, 2006.

[272] Hao Shao, TomÃ¡s Svoboda, Vittorio Ferrari, Tinne Tuytelaars, and Luc J. Van Gool, *Fast indexing for image retrieval based on local appearance with re-ranking.*, ICIP (3), 2003, pp. 737–740.

[273] Gaurav Sharma, Ritu Garg, and Santanu Chaudhury, *Curvature feature distribution based classification of indian scripts from document images*, Proceedings of the International Workshop on Multilingual OCR (New York, NY, USA), MOCR '09, ACM, 2009, pp. 3:1–3:6.

[274] Sheikh Faisal Rashid and Faisal Shafait and Thomas Breuel , *Connected Component level Multiscript Identification from Ancient Document Images*, Proceedings of the 9^{th} IAPR Workshop on Document Analysis System, 2010, pp. 1–4.

[275] R. Shekhar and C.V. Jawahar, *Word image retrieval using bag of visual words*, Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on, 2012, pp. 297–301.

[276] Haiying Shen, Ting Li, and Tom Schweiger, *An efficient similarity searching scheme in massive databases*, Proceedings of the 3^{th} International Conference on Digital Telecommunications (2008), 47–52.

[277] Jialie Shen, John Shepherd, and Anne H. H. Ngu, *Towards effective content-based music retrieval with multiple acoustic feature combination*, IEEE Transactions on Multimedia **8** (2006), no. 6, 1179–1189.

[278] Zhiyong Shen, Ping Luo, Yuhong Xiong, Jun Sun, and Yidong Shen, *Topic modeling for sequences of temporal activities*, Proceedings of the International Conference on Data Mining, 2009, pp. 980–985.

[279] Steven K. Shevell, *The science of color*, Elsevier Science & Technology, July 2003.

[280] P. Shivakumara, Trung Quy Phan, and C.L. Tan, *A laplacian approach to multi-oriented text detection in video*, Pattern Analysis and Machine Intelligence, IEEE Transactions on **33** (2011), no. 2, 412–419.

[281] P. Shivakumara, R.P. Sreedhar, Trung Quy Phan, Shijian Lu, and C.L. Tan, *Multi-oriented video scene text detection through bayesian classification and boundary growing*, Circuits and Systems for Video Technology, IEEE Transactions on **22** (2012), no. 8, 1227–1235.

[282] B. Siddiquie, R. S. Feris, and L. S. Davis, *Image ranking and retrieval based on multi-attribute queries*, Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (Washington, DC, USA), CVPR '11, IEEE Computer Society, 2011, pp. 801–808.

[283] P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, H. Meinedo, M. Bugalho, and I. Trancoso, *Temporal video segmentation to scenes using high-level audiovisual features*, IEEE Trans. Cir. and Sys. for Video Technol. **21** (2011), no. 8, 1163–1177.

[284] C. Silpa-Anan and R. Hartley, *Optimised kd-trees for fast image descriptor matching*, Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, 2008, pp. 1–8.

[285] Suranjit Sinha, Umapada Pal, and B. B. Chaudhuri, *Word-wise script identification from indian documents*, Document Analysis Systems, 2004, pp. 310–321.

[286] J. Sivic and A. Zisserman, *Video Google: A text retrieval approach to object matching in videos*, Proceedings of the International Conference on Computer Vision, vol. 2, October 2003, pp. 1470–1477.

[287] Alan F. Smeaton, Paul Over, and Aiden R. Doherty, *Video shot boundary detection: Seven years of trecvid activity*, Comput. Vis. Image Underst. **114** (2010), no. 4, 411–418.

[288] Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain, *Content-based image retrieval at the end of early years*, IEEE Transactions on Pattern Analysis and Machine Intelligence **22** (2000), no. 12, 1349–1380.

[289] J. Smith and S.-F. Chang, *Quad-tree segmentation for texture-based image query*, Proceedings of the second ACM international conference on Multimedia, 1994, pp. 279–286.

[290] John R. Smith and Shih-Fu Chang, *Visualseek: a fully automated content-based image query system*, Proceedings of the fourth ACM international conference on Multimedia, 1996, pp. 87–98.

[291] Cees G. M. Snoek and Marcel Worring, *Concept-based video retrieval*, Foundation and Trends of Information Retrieval **2** (2009), no. 4, 215–322.

[292] Yan Song, Yan-Tao Zheng, Sheng Tang, Xiangdong Zhou, Yongdong Zhang, Shouxun Lin, and Tat-Seng Chua, *Localized multiple kernel learning for realistic human action recognition in videos*, IEEE Transactions on Circuits Systems and Video Technology **21** (2011), no. 9, 1193–1202.

[293] Soren Sonneburg, Gunnar Ratsch, Christin Schafer, and Bernhard Scholkopf, *Large scale multiple kernel learning*, Journal of Machine Learning Research **7** (2006), 1531–1565.

[294] A. Lawrence Spitz, *Determination of the script and language content of document images*, IEEE Transactions on Pattern Analysis and Machine Intelligence **19** (1997), 235–245.

[295] C. Strouthopoulos, N. Papamarkos, A. Atsalakis, and C. Chamzas, *Text identification in color documents*, Proceedings of the 3^{rd} International Symposium on Image and Signal Processing and Analysis, 2003, pp. 702–705.

[296] Quan-Sen Sun, Sheng-Gen Zeng, Yan Liu, Pheng-Ann Heng, and De-Shen Xia, *A new method of feature fusion and its application in image recognition*, Pattern Recognition **38** (2005), no. 12, 2437 – 2448.

[297] Xinghua Sun, Mingyu Chen, and A. Hauptmann, *Action recognition via local descriptors and holistic features*, Proceedings of IEEE Computer Vision and Pattern Recognition Workshops, June 2009, pp. 58 –65.

[298] Shamik Sural and P. K. Das, *An mlp using hough transform based fuzzy feature extraction for bengali script recognition*, Pattern Recognition Letters **20** (1999), no. 8, 771 – 782.

[299] Kazem Taghva, Julie Borsack, and Allen Condit, *Effects of ocr errors on ranking and feedback using the vector space model*, International Journal of Information Processing and Management **32** (1996), 317–327.

[300] Atsuhiro Takasu, *Cross-lingual keyword recommendation using latent topics*, Proceedings of the 1^{st} International Workshop on Information Heterogeneity and Fusion in Recommender Systems (2010), 52–56.

[301] T. N. Tan, *Rotation invariant texture features and their use in automatic script identification*, IEEE Transactions on Pattern Analysis and Machine Intelligence **20** (1998), 751–756.

[302] Available at `http://www.cis.temple.edu/latecki/TestData/mpeg7shapeB.tar.gz` **MPEG-7 CE Shape-1: Part B**.

[303] *FlexiCapture 9.0* Available at `http://www.abbyy.com/`.

[304] *Oracle WebCenter Forms Recognition* Available at `http://www.oracle.com/`.

[305] *Teleform* Available at `http://www.caylx.com.au/`.

[306] `http://lucene.apache.org/`.

[307] `http://lucene.apache.org/core/4_0_0-BETA/core/org/apache/lucene/util/automaton/package-summary.html`.

[308] Micheal E. Tipping, *The relevance vector machine*, Proceedings of Advances in Neural Information Processing Systems 12, MIT Press, 2000, pp. 652–658.

[309] D. Tuia, G. Camps-Valls, G. Matasci, and M. Kanevski, *Learning Relevant Image Features With Multiple-Kernel Classification*, IEEE Transactions on Geoscience and Remote Sensing, **48** (2010), no. 10, 3780–3791.

[310] R. Sinan Tumen, M. Emre Acer, and T. Metin Sezgin, *Feature extraction and classifier combination for image-based sketch recognition*, Proceedings of the Seventh Sketch-Based Interfaces and Modeling Symposium (Aire-la-Ville, Switzerland, Switzerland), SBIM '10, Eurographics Association, 2010, pp. 63–70.

[311] Athitsos Vassilis, Potamias Michalis, Papapetrou Panagiotis, and Kollios George, *Nearest neighbor retrieval using distance based hashing*, Proceedings of the 24^{th} International Conference on Data Engineering, April 2008, pp. 327–336.

[312] Andrea Vedaldi, Varun Gulshan, Manik Varma, and Andrew Zisserman, *Multiple kernels for object detection*, Proceedings of the 12^{th} IEEE International Conference on Computer Vision, 2009, pp. 606–613.

[313] S. Vijayanarasimhan and K. Grauman, *Top-down pairwise potentials for piecing together multi-class segmentation puzzles*, Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, june 2010, pp. 25 –32.

[314] Paul Viola and Michael J. Jones, *Robust real-time face detection*, International Journal of Computer Vision **57** (2004), 137–154.

[315] Julia Vogel and Bernt Schiele, *Semantic modeling of natural scenes for content-based image retrieval*, Int. J. Comput. Vision **72** (2007), no. 2, 133–157.

[316] Daniel D. Walker, William B. Lund, and Eric K. Ringger, *Evaluating models of latent document semantics in the presence of ocr errors*, Proceeding the 2010 Conference on Empirical Methods in Natural Language Processing, 2010, pp. 240–250.

[317] Changhu Wang, Feng Jing, Lei Zhang, and Hong-Jiang Zhang, *Scalable search-based image annotation of personal images*, Proceedings of the 8th ACM international workshop on Multimedia information retrieval, MIR '06, 2006, pp. 269–278.

[318] Dingding Wang, Shenghuo Zhu, Tao Li, and Yihong Gong, *Multi-document summarization using sentence-based topic models*, Proceeding of the ACL-IJCNLP 2009 (2009), 297–300.

[319] Heng Wang, Muhammad Muneeb Ullah, Alexander Kläser, Ivan Laptev, and Cordelia Schmid, *Evaluation of local spatio-temporal features for action recognition*, Proceedings of the British Machine Vision Conference, 2009, p. 127.

[320] Jun Wang, Sanjiv Kumar, and Shih-Fu Chang, *Semi-supervised hashing for scalable image retrieval*, Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (2010), 3424–3431.

[321] ———, *Sequential projection learning for hashing with compact codes*, Proceedings of International Conference on Machine Learning, 2010, pp. 1127–1134.

[322] Kai Wang, B. Babenko, and S. Belongie, *End-to-end scene text recognition*, Computer Vision (ICCV), 2011 IEEE International Conference on, 2011, pp. 1457–1464.

[323] Tao Wang, Jianguo Li, Qian Diao, Wei Hu, Yimin Zhang, and C. Dulong, *Semantic event detection using conditional random fields*, Proceedings of the Computer Vision and Pattern Recognition Workshop, 2006, p. 109.

[324] Yong Wang, Tao Mei, Shaogang Gong, and Xian-Sheng Hua, *Combining global, regional and contextual features for automatic image annotation*, Pattern Recognition **42** (2009), 259–266.

[325] Toyohide Watanabe and Tsuneo Sobue, *Layout analysis of complex documents*, Proceedings of the 15^{th} International Conference on Pattern Recognition, vol. 4, 2000, pp. 447 – 450.

[326] Shikui Wei, Yao Zhao, Zhenfeng Zhu, and Nan Liu, *Multimodal fusion for video search reranking*, IEEE Trans. on Knowl. and Data Eng. **22** (2010), no. 8, 1191–1199.

[327] Wang Weihong and Wang Song, *A scalable content-based image retrieval scheme using locality-sensitive hashing*, Proceedings of the International Conference on Computational Intelligence and Natural Computing **1** (2009), 151–154.

[328] Liu Wenyin, Yanfeng Sun, and Hongjiang Zhang, *Mialbum - a system for home photo managemet using the semi-automatic image annotation approach*, Proceedings of the 8^{th} ACM international conference on Multimedia, MULTIMEDIA '00, ACM, 2000, pp. 479–480.

[329] Jason Weston, Samy Bengio, and Nicolas Usunier, *Large scale image annotation: learning to rank with joint word-image embeddings*, Mach. Learn. **81** (2010), no. 1, 21–35.

[330] David A. White and Ramesh Jain, *Similarity indexing: Algorithms and performance*, In Storage and Retrieval for Image and Video Databases (SPIE, 1996, pp. 62–73.

[331] Michael L. Wick, Michael G. Ross, and Erik G. Learned-Miller, *Context-sensitive error correction: Using topic models to improve ocr*, Proceedings of the 9^{th} International Conference on Document Analysis and Recognition (2007), 1168–1172.

[332] Geert Willems, Tinne Tuytelaars, and Luc Gool, *An efficient dense and scale-invariant spatio-temporal interest point detector*, Proceedings of the 10^{th} European Conference on Computer Vision: Part II, 2008, pp. 650–663.

[333] Erling Wold, Thom Blum, Douglas Keislar, and James Wheaton, *Content-based classification, search, and retrieval of audio*, IEEE MultiMedia **3** (1996), no. 3, 27–36.

[334] Wing Seong Wong, Nasser Sherkat, and Tony Allen, *Use of colour in form layout analysis*, Proceedings of the 6^{th} International Conference on Document Analysis and Recognition, 2001, pp. 942 – 946.

[335] S.L. Wood, Xiaozhong Yao, K. Krishnamurthi, and L. Dang, *Language identification for printed text independent of segmentation*, International Conference on Image Processing, vol. 3, october 1995, pp. 428 –431.

[336] S. Wshah, G. Kumar, and V. Govindaraju, *Script independent word spotting in offline handwritten documents based on hidden markov models*, Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on, 2012, pp. 14–19.

[337] Lei Wu, Rong Jin, and A.K. Jain, *Tag completion for image retrieval*, Pattern Analysis and Machine Intelligence, IEEE Transactions on **35** (2013), no. 3, 716–727.

[338] V. Wu, R. Manmatha, and E.M. Riseman, *Textfinder: an automatic system to detect and recognize text in images*, Pattern Analysis and Machine Intelligence, IEEE Transactions on **21** (1999), no. 11, 1224–1229.

[339] Zhong Wu, Qifa Ke, Michael Isard, and Jian Sun, *Bundling features for large scale partial-duplicate web image search*, CVPR'09, 2009, pp. 25–32.

[340] Changsheng Xu, Jinjun Wang, Kongwah Wan, Yiqun Li, and Lingyu Duan, *Live sports event detection based on broadcast video and web-casting text*, Proceedings of the 14^{th} annual ACM international conference on Multimedia, MULTIMEDIA '06, 2006, pp. 221–230.

[341] Changsheng Xu, Yi-Fan Zhang, Guangyu Zhu, Yong Rui, Hanqing Lu, and Qingming Huang, *Using webcast text for semantic event detection in broadcast sports video*, Multimedia, IEEE Transactions on **10** (2008), no. 7, 1342 –1355.

[342] Jun Xu and Hang Li, *Adarank: a boosting algorithm for information retrieval*, Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '07, 2007, pp. 391–398.

[343] Antonio T. Yair Weiss, *Spectral hashing*, Proceedings of NIPS, 2008, pp. 1753–1760.

[344] Oksana Yakhnenko and Vasant Honavar, *Annotating images and image objects using a hierarchical dirichlet process model*, Proceedings of the 9^{th} International Workshop on Multimedia Data Mining, 2008, pp. 1–7.

[345] Akira Yanagawa, Shih-Fu Chang, Lyndon Kennedy, and Winston Hsu, *Columbia university's baseline detectors for 374 lscom semantic visual concepts*, Tech. report, Columbia University, March 2007.

[346] Changbo Yang, Ming Dong, and Farshad Fotouhi, *Region based image annotation through multiple-instance learning*, Proceedings of 13^{th} annual ACM international conference on Multimedia, 2005, pp. 435–438.

[347] Liu Yang and Rong Jin, *Distance Metric Learning: A Comprehensive Survey*, Tech. report, Department of Computer Science and Engineering, Michigan State University.

[348] Qixiang Ye, Qingming Huang, Wen Gao, and Debin Zhao, *Fast and robust text detection in images and video frames*, Image and Vision Computing **23** (2005), no. 6, 565 – 576.

[349] Jie Yin, Derek Hao Hu, and Qiang Yang, *Spatio-temporal event detection using dynamic conditional random fields*, Proceedings of the 21^{st} International Joint Conference on Artificial Intelligence, 2009, pp. 1321–1326.

[350] Yossi Zana and Roberto M. Cesar, Jr, *Face recognition based on polar frequency features*, ACM Trans. Appl. Percept. **3** (2006), no. 1, 62–82.

[351] Lei Zhang, Jingxin Chang, Xuezhi Xiang, and Xiaosen Feng, *Topic indexing of spoken documents based on optimized n-best approach*, Proceedings of the International Conference on Intelligent Computing and Intelligent Systems, 2009.

[352] Lei Zhang and Jun Ma, *Image annotation by incorporating word correlations into multi-class svm*, Soft Computing **15** (2011), no. 5, 917–927.

[353] Wenchao Zhang, Shiguang Shan, Wen Gao, Yizheng Chang, Bo Cao, and Peng Yang, *Information fusion in face identification*, Proceedings of the 17^{th} International Conference on Pattern Recognition, vol. 3, 2004, pp. 950–953.

[354] Zhi-Hua Zhou, Ke-Jia Chen, and Hong-Bin Dai, *Enhancing relevance feedback in image retrieval using unlabeled data*, ACM Trans. Inf. Syst. **24** (2006), no. 2, 219–244.

[355] Guangyu Zhu, Ming Yang, Kai Yu, Wei Xu, and Yihong Gong, *Detecting video events based on action recognition in complex scenes using spatio-temporal descriptor*, Proceedings of the 17^{th} ACM international conference on Multimedia, 2009, pp. 165–174.

[356] Guangyu Zhu, Yefeng Zheng, David Doermann, and Stefan Jaeger, *Signature detection and matching for document image retrieval*, IEEE Transactions on Pattern Analysis and Machine Intelligence **31** (2009), no. 11, 2015–2031.

[357] Yue-Ting Zhuang, Yi Yang, and Fei Wu, *Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval*, IEEE Transactions on Multimedia **10** (2008), no. 2, 221 – 229.

[358] M. Zimmermann, J.-C. Chappelier, and H. Bunke, *Offline grammar-based recognition of handwritten sentences*, Pattern Analysis and Machine Intelligence, IEEE Transactions on **28** (2006), no. 5, 818–821.

[359] A. Zolnay, R. Schlueter, and H. Ney, *Acoustic feature combination for robust speech recognition*, Proceedings of International Conference on Acoustics, Speech, and Signal Processing, 2005., vol. 1, 2005, pp. 457 – 460.

Appendix A

Locality Sensitive Hashing

Locality Sensitive Hashing (LSH) provides a probabilistic solution for the approximate nearest neighbour search problem [141]. For solving (c, R) nearest neighbour search problem in the space $\mathcal{R}^d$ with defined distance measure d, an locality sensitive hash function family is defined as

- The family $\mathbb{H} = \{f : \mathcal{R}^d \rightarrow U\}$ of hashing functions is called (R, cR, P_1, P_2) sensitive if for points $p, q \in \mathcal{R}^d$
  - If $d(p, q) \leq R$ then $[f(p) = f(q)] \geq P_1$
  - If $d(p, q) \geq cR$ then $[f(p) = f(q)] \leq P_2$

Here c is a real number greater than 1.

The LSH function family would be useful, if it satisfies the inequality $P_1 \geq P_2$. Therefore, if point q is close to p, the hash value of p and q would be same, i.e., both points would be hashed in same bucket. Whereas if q is placed far from p, then p and q would be

less likely to be hashed in the same bucket. The difference between P_1 and P_2 can further be increased by performing projection by combination of k functions selected randomly, i.e., formulating hashing function $g(\cdot)$ by concatenation of k hashing functions. Since $(P_1/P_2)^k > (P_1/P_2)$, it increases the ratio of probabilities of separation. The k-bit real number obtained for each data point after projection is the corresponding hash index. The nearest neighbour search for the query point can be performed by mapping the query to hash space using k-bit function and then performing a linear search over the points falling into the same bucket as query. The search success rate in any projection is increased by generating multiple hash tables, i.e., L independent hash tables, and collecting the neighbours from these tables to find the nearest neighbour. The large value of parameter k increases the precision of the retrieval, however the recall rate decreases because of the exponential decrease in collision probability. To ensure satisfactory recall multiple hash tables are required. Large number of hash tables, i.e., large L increases the recall at increased search complexity, but simultaneously precision decreases. Therefore, the selection of L and k should be optimal to distribute the data points sparsely to maintain the advantage of approximate nearest neighbour search.

Appendix B

Relevance Vector Machine for Classification

Relevance Vector Machine (RVM) is general Bayesian framework for obtaining sparse solutions to regression and classification utilizing models linear in parameters [308]. RVM has identical functional form to the popular and state-of-the-art 'support vector machine' (SVM).

We consider two class binary problem with binary target variable $t \in \{0, 1\}$. The model can be expressed as linear combination of basis functions transformed by a logistic sigmoid function

$$y(\mathbf{x}, \mathbf{w}) = \sigma(\mathbf{w}^T \phi(\mathbf{x})) \tag{B.0.1}$$

Where $\sigma(.)$ is the logistic function defined as $\sigma(y) = 1/(1+e^{-y})$. If we introduce gaussian

prior over the weight vector **w**, then we obtain

$$p(\mathbf{t}|\mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n} \{1 - y_n\}^{1-t_n} \tag{B.0.2}$$

Here we use ARD prior where there is a separate predision hyper-parameter associated with each weight parameter i.e.

$$p(\mathbf{w}|\alpha) = \prod_{i=1}^{M} N(w_i|0, \alpha_i^{-1}) \tag{B.0.3}$$

where α_i represents the precision of the corresponding parameter w_i, and α denotes the $(\alpha_1, ..., \alpha_M)^T$. We follow Laplace approximation to integrate over **w**.

Laplace approximation aims to find a Gaussian approximation to a probability density defined over a set of continuous variables such that Gaussian approximation is centred on a mode of the distribution. The first step is to find the mode of the distribution.

For a fixed value of α the mode of the posterior distribution over **w** nis obtained by maximizing

$$\begin{aligned} \ln p(\mathbf{w}|\mathbf{t}, \alpha) &= \ln \{p(\mathbf{t}|\mathbf{w})p(\mathbf{w}|\alpha)\} - \ln p(\mathbf{t}|\alpha) \\ &= \sum_{n=1}^{N} \{t_n \ln y_n + (1 - t_n)\ln(1 - y_n)\} - \frac{1}{2}\mathbf{w}^T\mathbf{A}\mathbf{w} + const \end{aligned} \tag{B.0.4}$$

Where **A** = diag(α_i). The mode can be calculated by Iterative Reweighted Least Squares(IRLS).

We need gradient vector and hessian matrix of the log posterior distribution.

$$\nabla \ln p(\mathbf{w}|\mathbf{t}, \boldsymbol{\alpha}) = \mathbf{\Phi}^T(\mathbf{t} - \mathbf{y}) - \mathbf{A}\mathbf{w} \tag{B.0.5}$$

$$\nabla\nabla \ln p(\mathbf{w}|\mathbf{t}, \boldsymbol{\alpha}) = -(\mathbf{\Phi}^T\mathbf{B}\mathbf{\Phi} + \mathbf{A}) \tag{B.0.6}$$

where **B** is $N \times N$ diagonal matrix with elements $b_n = y_n(1 - y_n)$, the vector $\mathbf{y} = (y_1, ..., y_N)^{\mathrm{T}}$, and $\mathbf{\Phi}$ is the design matrix with elements $\Phi_{ni} = \phi_i(\mathrm{x}_n)$. At the convergence

of the IRLS algorithm, the negative of the Hessian represents the inverse covariance matrix for the Gaussian approximation to the posterior distribution.

The mode of the resulting approximation to the posterior distribution, corresponds to the mean of the Gaussian distribution,

$$\mathbf{w}^* = \mathbf{A}^{-1}\mathbf{\Phi}^T(\mathbf{t} - \mathbf{y}) \tag{B.0.7}$$

$$\Sigma = (\mathbf{\Phi}^{\mathrm{T}}\mathbf{B}\mathbf{\Phi} + \mathbf{A})^{-1} \tag{B.0.8}$$

Given a new test point x_*, predictions are made for the corresponding target t_*, in terms of predictive distribution:

$$p(t_*|\mathbf{t}) = \int p(t_*|\mathbf{w}, \boldsymbol{\alpha})p(\mathbf{w}, \boldsymbol{\alpha}|\mathbf{t})d\mathbf{w}d\boldsymbol{\alpha} \tag{B.0.9}$$

Evaualtion of $p(\mathbf{w}, \boldsymbol{\alpha}|\mathbf{t})$ si directly not possible, since we cannot perform the normalising integral. We decompose the posterior as:

$$p(\mathbf{w}, \boldsymbol{\alpha}|\mathbf{t}) = p(\mathbf{w}|\boldsymbol{\alpha}, \mathbf{t})p(\boldsymbol{\alpha}|\mathbf{t}) \tag{B.0.10}$$

Further $p(\boldsymbol{\alpha}|\mathbf{t}) \propto p(\mathbf{t}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})$, Where $p(\boldsymbol{\alpha}) = \prod_{i=0}^{N} \text{Gamma}(\alpha_i|a, b)$. To make these priors non informatiove we might fix $a = b = 10^{-4}$. Setting these parameters to zero, we obtain uniform hyperpriors, which gives scale invariance. Predictions are invariant of the scale of both **t** and the basis function outputs both. This formulation of prior distribution is a type of *Automatic Relevance Determination*. Using broad prior over the hyper-parameters allows the posterior probability mas to concentrate at very large values of some of these α variables, with the consequence that the posterior probability of the associated weights will be concentrated at zero, thus effectively 'switching off' the corresponding inputs, and

so deeming them to be irrelevant. Since already have $\mathbf{w}^*$ and Σ we can evaluate margnila likelihood as

$$\begin{aligned} p(\mathbf{t}|\boldsymbol{\alpha}) &= \int p(\mathbf{t}|\mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha})d\mathbf{w} \\ &\cong p(\mathbf{t}|\mathbf{w}^*)p(\mathbf{w}^*|\boldsymbol{\alpha})(2\pi)^{M/2}|\Sigma|^{1/2} \end{aligned} \tag{B.0.11}$$

Equationg the derivative of the marginal likelihood with respect to α_i equalt to zero, we obtain

$$-\frac{1}{2}(w_i^*)^2 + \frac{1}{2\alpha_i} - \frac{1}{2}\Sigma_{ii} = 0 \tag{B.0.12}$$

Defining $\gamma_i = 1 - \alpha_i\Sigma_{ii}$ we get

$$\alpha_i^{new} = \frac{\gamma_i}{(w_i^*)^2} \tag{B.0.13}$$

The above equation is the hyperparameter reestimation formula which can be used for prediction after convergence. In the case of *K* class problem we have *K* linear models of the form

$$a_k = \mathbf{w}_k^T\mathbf{x} \tag{B.0.14}$$

which are combined using a softmax function to give outputs

$$y_k(\mathbf{x}) = \frac{exp(a_k)}{\sum_j exp(a_j)} \tag{B.0.15}$$

the log likelihood function is given by

$$\ln p(\mathbf{T}|\mathbf{w}_1, ..., \mathbf{w}_K) = \prod_{n=1}^{N}\prod_{k=1}^{K} y_{nk}^{t_{nk}} \tag{B.0.16}$$

Where t_{nk} have 1-of-*K* coding of each datapoint n, and **T** is a mtrix with elements t_{nk}. The principal disadvantage is that hessian matrix has size $MK \times MK$, where *M* is the number

of the active basis functions, which gives an additional factor of K^3 in the computational cost of training compared to the two-class RVM.

Appendix C

Conditional Random Fields

Conditional Random Fields (CRFs) are discriminative modelling algorithm for segmenting and labelling the sequential data [164]. It is an effective approach for supervised structure learning of the relationship between complex objects such as graphs. Consider $\mathbf{X}$ as random variable over the data sequence, and $\mathbf{Y}$ as random variable over corresponding label sequences. If $\mathbf{Y}$ is indexed over the graph G with edges V and nodes E and $\mathbf{Y} = (\mathbf{Y}_v)_{v \in V}$. Then the $(\mathbf{X}, \mathbf{Y})$ defines a Conditional Random Field by conditioning the random variable $\mathbf{Y}_v$ on $\mathbf{X}$ such that $\mathbf{Y}_v$ follows Markov property in G as $P(\mathbf{Y}_v|\mathbf{X}, \mathbf{Y}_w, w \neq v) = P(\mathbf{Y}_v|\mathbf{X}, \mathbf{Y}_w, w \sim v)$. $\mathbf{X}$ may have any graph structure irrespective of the structure of $\mathbf{Y}$. The joint distribution over $\mathbf{Y}$ has following form

$$p(\mathbf{y}|\mathbf{x}) \propto \exp(\sum_{e \in E,k} \lambda_k f_k(e, \mathbf{y}|_e, \mathbf{x}) + \sum_{v \in V,k} \mu_k g_k(v, \mathbf{y}|_v, \mathbf{x})) \tag{C.0.1}$$

Here, $\mathbf{x}$ and $\mathbf{y}$ are data and label sequence and $\mathbf{y}|_S$ is set of components of $\mathbf{y}$ associated with the vertices in sub-graph S. Function f_k defines the input dependent evidences and

g_k represents the pair-wise relationship between labels of the sequence data, and λ_k and μ_k are the associated Lagrange parameters. The parameter estimation problem determines the parameters $\theta = (\lambda_1, \mu_1, \lambda_2, \mu_2, ...)$ by the maximization of log-likelihood objective as

$$O(\theta) = \sum_{i=1}^{N} log\ p_\theta(\mathbf{y}^i|\mathbf{x}^i) \propto \sum_{\mathbf{x},\mathbf{y}} p'(\mathbf{x}, \mathbf{y}) log\ p_\theta(\mathbf{y}|\mathbf{x}) \tag{C.0.2}$$

$p'(\mathbf{x}, \mathbf{y})$ is the empirical probability distribution which generated training set. The parameter estimation for simple linear chain CRF using iterative scaling is discussed in [164]. Recent development in this direction have presented gradient descent based methods such as Newton and Quasi-Newton methods, stochastic gradient methods and stochastic meta descent methods.

Appendix D

Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) defines a generative probabilistic model over the collection of documents [31]. The plate diagram for LDA generative process is shown in the figure D.1. The outer plate represents the documents and inner plate represents the topic sampling over set of words. M denotes number of documents in the collection and N represents a number of words in a document. In the context of topic modelling, docu-

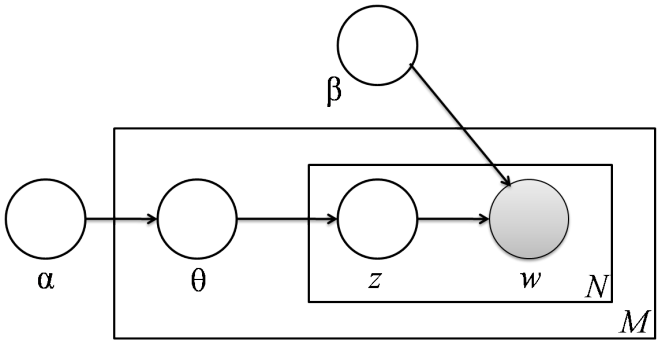


Figure D.1: Graphical model for LDA

ments can be text corpus or image collections. The documents are represented as random

mixtures over latent topics, where each topic is characterized by a distribution over words. In case of images, local properties are defined as terms. The documents are represented as as random mixtures over latent topics, where each topic is characterized by a distribution over words. The generative process for each document **w** in the collection D is defined as follows:

- Select N = $Poisson(\zeta)$
- Choose θ = $Dirichlet(\alpha)$
- For each word w_n of the document **w**: select a topic z_n = $Multinomial(\theta)$, select w_n from multinomial probability $p(w_n|z_n;\beta)$

Parameter β in the figure D.1 represents topic/word probabilities as fixed quantities which needs to be estimated. However for most of the application smooth LDA model is applied which explicitly models β as random variable. The inference step includes the computation of posterior distribution of the hidden variables for a given document.

$$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)} \tag{D.0.1}$$

The joint distribution in the numerator of equation (D.0.1) is defined as

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha)\prod_{i=1}^{N} p(z_i|\theta)p(w_i|z_i, \beta)$$

The marginal distribution $p(\mathbf{w}|\alpha, \beta)$ is computed as follows

$$p(\mathbf{w}|\alpha, \beta) = \int p(\theta|\alpha)\left(\prod_{i=1}^{N}\sum_{z_i} p(z_i|\theta)p(w_i|z_i, \beta)\right)d\theta \tag{D.0.2}$$

Computing the marginal probability over document collection D as

$$p(D|\alpha, \beta) = \prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{i=1}^{N_d}\sum_{z_{di}} p(z_d i|\theta_d)p(w_{di}|z_{di}, \beta)\right)d\theta_d$$

The intractable form of equation (D.0.2) leaves (D.0.1) exact inferencing unachievable. The problems is solved by applying approximate inference methods (Viz. Laplace approximation, Variational approximation, and Markov chain monte carlo simulation) for inferencing and parameter estimation. In this context, a simple convexity based variational inferencing method is presented in [31].

Publications

Journal

1. Ehtesham Hassan, Santanu Chaudhury, M Gopal, “Feature Combination in Kernel Space for Distance based Image Hashing”, *Accepted for publication in IEEE Transactions on Multimedia (IEEE TMM)*.

2. Ehtesham Hassan, Santanu Chaudhury, M Gopal, “Word Shape Descriptor Based Document Image Indexing: A New DBH Based Approach”, *Accepted for publication in the International Journal of Document Analysis and Recognition (IJDAR)*.

Conference

1. Ehtesham Hassan, Santanu Chaudhury, M Gopal, Vikram Garg, “A Hybrid Framework for Event Detection Using Multi-modal Features", *In the Proceedings of 3rd International Workshop on Video Event Categorization, Tagging and Retrieval for Real -World Applications: ICCV-2011*, on 13^{th} November, Barcilona, Spain, pp. 1510-1515, 2011.

2. Ehtesham Hassan, Santanu Chaudhury, M Gopal, “Annotating Dance Posture Images Using Multi Kernel Feature Combination", *In the Proceedings of National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics*, on 15^{th} - 17^{th} December, Hubli, India, 2011.

3. Ehtesham Hassan, Santanu Chaudhury, M Gopal, “Document Image Indexing using Edit Distance based Hashing", *In the Proceedings of 11th International Conference on Document Analysis and Recognition*, on 18^{th} - 21^{st} September, Beijing, China, pp. 1200-1206, 2011.

4. Ehtesham Hassan, Vikram Garg, S. K. Mirajul Haque, Santanu Chaudhury, M Gopal, “Searching OCRed text: An LDA based Approach", *In the Proceedings of 11th International Conference on Document Analysis and Recognition*, on 18^{th} - 21^{st} September, Beijing, China, pp. 1210-1214, 2011.

5. Ritu Garg, Ehtesham Hassan, Santanu Chaudhury, M Gopal, “A CRF Based Scheme for Overlapping Multi-Colored Text Graphics Separation", *In the Proceedings of 11th International Conference on Document Analysis and Recognition*, on 18^{th} - 21^{st} September, Beijing, China, pp. 1215-1219, 2011.

6. Ehtesham Hassan, Ritu Garg, Santanu Chaudhury, M Gopal, “Script based Text Identification: A Multi-level Architecture", *In the Proceedings of Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data*, on 17^{th} September, Beijing, China, 2011.

7. Ehtesham Hassan, Santanu Chaudhury, M Gopal, “Multiple Kernel Learning for Image Indexing", *In the Proceedings of* 7^{th} *Indian Conference on Computer Vision, Graphics and Image Processing*, 12^{th} - 15^{th} December, Chennai, India, pp. 353-358, 2010.

8. Ehtesham Hassan, Santanu Chaudhury, M Gopal, “Document Image Retrieval Using Feature Combination in Kernel Space", *In the Proceedings of International Conference on Pattern Recognition*, 23^{th} - 26^{th} August, Istanbul, Turkey, pp. 2009-2012, 2010.

9. Ehtesham Hassan, Santanu Chaudhury, M Gopal and Jignesh Dholakia, “Use of MKL as Symbol Classifier for Gujarati Character Recognition", *In the Proceedings of* 9^{th} *International Workshop on Document Analysis Systems*, 9^{th} - 11^{th} June, Boston, USA, pp. 255-262, 2010.

10. Ehtesham Hassan, Santanu Chaudhury, and M Gopal, “Shape Descriptor based Document Image Indexing and Symbol Recognition", *In the Proceedings of 10th International Conference on Document Analysis and Recognition*, 26^{th} - 29^{th} July, Barcelona, pp. 206-210, 2009.

Biography

Ehtesham Hassan was born in Jaunpur, India in 1982. He received M.Tech in Electronics and Communication Engineering from Indian Institute ofTechnology Roorkee in 2005, and B.E. degree in Electrical and Electronics Engineering from Visvesvaraya Technological University Belgaum, India in 2002. He worked for Samsung India Software Operations, Bangalore as senior software engineer from July 2005 to June 2007. His research interests include multimedia analysis, pattern recognition and machine learning.