IMAGE RETRIEVAL USING VISUAL IMAGE FEATURES AND AUTOMATIC IMAGE ANNOTATION

Nouman Ali

2011-UET/PhD-CP-23

Thesis Supervisor

Dr. Khalid Bashir Bajwa
Assistant Professor

Department of Computer Engineering,
University of Engineering and Technology, Taxila,
Pakistan.

August-2016
IMAGE RETRIEVAL USING VISUAL IMAGE FEATURES AND AUTOMATIC IMAGE ANNOTATION

Nouman Ali

2011-PhD/UET-CP-23

A thesis submitted in partial fulfillment of the requirement for the

Degree of Doctor of Philosophy

Thesis Supervisor

Dr. Khalid Bashir Bajwa

Assistant Professor

Department of Computer Engineering,
University of Engineering and Technology, Taxila,
Pakistan.

August-2016
DECLARATION

I undertake that all the material presented in this thesis is my own work and is not written for me, in whole or in parts by any other person. I also undertake that any quotations or paraphrases from the published work of another person are duly acknowledged and cited in this thesis.

Nouman Ali

2011-PhD/UET-CP-23
Plagiarism Undertaking

I take full responsibility of the research work conducted during the PhD Thesis titled Image Retrieval Using Visual Image Features and Automatic Image Annotation. I solemnly declare that the research work presented in the thesis is done solely by me with no significant help from any other person. I have also written the complete thesis by myself. Moreover, I have not presented this thesis (or substantially similar research work) or any part of the thesis previously to any other degree awarding Institution in Pakistan or abroad.

Therefore, I as an author of the above-mentioned thesis, solemnly declare that no portion of my thesis has been plagiarized and the material used in this thesis from other sources is properly referred.

Nouman Ali

2011-PhD/UE-CP-23
ACKNOWLEDGEMENT

First of all, I am thankful to Allah Almighty for allowing me to complete PhD thesis in time. I am thankful to my supervisor Dr. Khalid Bashir Bajwa for his supervision and guidance throughout the development of this thesis and for providing academic opportunities and intellectual freedom to accomplish this research.

Additionally I would like to thank the members of my Research Monitoring Committee (Dr. Adeel Akram, Dr. Iram Biag, Dr. Zulfiqar Hassan Khan, Dr. Tabassam Nawaz and Dr. Salabat Khan) for their valuable comments and feedback that improved the quality of this thesis.

I am thankful to the Higher Education Commission (HEC) of Pakistan for the fellowship grant, which made it possible for the research to be carried out at the Institute of Computer Aided Automation, Computer Vision Lab, Vienna University of Technology, Austria. Prof. Robert Sablatnig hosted my research visit and provided an access to state-of-the-art MATLAB servers of Vienna University of Technology, Austria for the simulation of this research. Dr. Hafeez Anwar (Assistant Professor, Comsats Attock) provided his valuable guideline during my stay at Computer Vision Lab, Vienna University of Technology, Austria. Prof. Oge Marques (Florida Atlanta University, USA) and Prof. Dr Jonathan Loo (Middlesex University, London, UK) evaluated my thesis as foreign experts. Amy Bruno-Linder, University of Vienna, Austria, provided her valuable suggestions concerning the language use during the writing of research papers.

I would like to say thanks to all of my colleagues and friends including Dr. Zeshan Iqbal, Zakaullah and Jawad Ahmed who were a great source of encouragement for me since last five years.
DEDICATION

My Parents

&

My family members
ABSTRACT

In recent few years, the exponential growth in the number of multimedia databases makes Content-Based Image Retrieval (CBIR) a challenging research area. In image classification and retrieval problems, the extraction of a meaningful image descriptor is an active research area. In CBIR, feature vectors are used to represent the images that are commonly based on color, texture, shape and spatial layout of the image. The mid-level local feature detectors are applied to map the image representation in a high-dimension feature space. The Bag of Features (BoF) based image representation model is widely used for CBIR and local descriptors are commonly applied to extract the visual features. In BoF representation model, an image is represented as an order-less histogram while the spatial contents provides the discriminating details that are useful for image retrieval and classification-based problems.

The spatial information is added to the inverted index of BoF representation model by computing histograms of visual words over the triangular regions of the image. An Image is divided into two and four triangular regions that are referred as histograms of triangles Level 1 and Level 2, respectively. Image representation as triangular histograms is selected to extract the spatial contents from top, down, left and right that are in form of regions and objects of interest. The proposed image representation is evaluated by applying three state-of-the-art classifiers and two standard image benchmarks are selected to determine the best retrieval performance. According to the experimental results, Deep Belief Network (DBN) consisting of auto-encoders outperforms Support Vector Machines (SVM) and Radial Basis Function Neural Network (RBF-NN). In addition to this, multi-label Automatic Image Annotation (AIA) is used to describe the image in the form of high-level semantics.

The visual words integration (also known as late fusion) of Scale Invariant Feature Transform (SIFT) and Speeded-Up Robust Features (SURF) is selected to enhance the performance of image retrieval. The two local features representations are selected for image retrieval because SIFT is more robust to the change in scale and rotation, while SURF is robust to changes in illumination. The proposed visual words integration is evaluated on five standard image benchmarks. The visual words integration of SIFT and SURF is selected to enhance the effectiveness and reliability of image retrieval. Image representation based on the visual words integration of SIFT and SURF adds the robustness of both local features for an effective and reliable image retrieval.
# TABLE OF CONTENTS

ABSTRACT ...................................................................................................................... i
TABLE OF CONTENTS ................................................................................................... ii
LIST OF FIGURES ......................................................................................................... iv
LIST OF TABLES ........................................................................................................... vi
LIST OF ACRONYMS .................................................................................................. vii
LIST OF PUBLICATIONS ............................................................................................ ix

1 Introduction .............................................................................................................. 1
   1.1 Research Motivation ............................................................................................ 2
   1.2 Research Objectives ........................................................................................... 3
   1.3 Research Contributions ....................................................................................... 4
   1.4 Structure of the Thesis ....................................................................................... 4

2 Fundamentals of Image Retrieval ............................................................................ 6
   2.1 Basic Architecture of CBIR ............................................................................... 6
   2.2 Feature Extraction Techniques .......................................................................... 7
       2.2.1 Image segmentation ............................................................................... 7
       2.2.2 Color features ....................................................................................... 10
       2.2.3 Texture features .................................................................................... 14
       2.2.4 Interest point detectors ......................................................................... 19
       2.2.5 Shape features ..................................................................................... 21
       2.2.6 Combination of visual features ............................................................. 22

   2.3 Performance Evaluation .................................................................................. 24
       2.3.1 Precision and recall .............................................................................. 24
       2.3.2 F-measure ............................................................................................ 25
       2.3.3 Average-precision ............................................................................... 25
       2.3.4 Mean average precision ...................................................................... 25
       2.3.5 Precision-recall curve ......................................................................... 26

   2.4 Image Benchmarks/ Datasets .......................................................................... 26

   2.5 Bag of Features (BoF) Model .......................................................................... 26
       2.5.1 Extraction of local features .................................................................. 28
       2.5.2 Clustering .............................................................................................. 29
       2.5.3 Encoding of feature space ..................................................................... 29
       2.5.4 Inverted index for BoF model ............................................................... 29

   2.6 Image Representation Using BoF Model ........................................................ 29
       2.6.1 Image Classification ............................................................................. 30
       2.6.2 Distance measures .............................................................................. 30

   2.7 Types of Local Features .................................................................................. 31
       2.7.1 Interest point-detectors/Sparse features ............................................. 31
       2.7.2 Dense features extraction .................................................................... 31
       2.7.3 Random features extraction ................................................................. 32
       2.7.4 Dense interest point sampling .............................................................. 32

   2.8 Clustering ........................................................................................................... 33
References

Chapter 3

3 Histogram of Triangular Regions

3.1 Introduction

3.2 Bag of Features Based Image Representation

3.2.1 Computation of triangular histograms

3.3 Image Classification Framework

3.3.1 Support vector machines

3.3.2 Radial basis function neural networks

3.3.3 Deep belief networks

3.4 Experimental Parameters

3.4.1 Performance on the Corel A image benchmark

3.4.2 Performance on the Fifteen Scene image benchmark

3.4.3 Run Time Analysis

3.5 Chapter Summary

Chapter 4

4 Visual Words Integration of SIFT and SURF

4.1 Introduction

4.2 Visual Words Integration of SIFT-SURF

4.2.1 Scale invariant feature transform (SIFT)

4.2.2 Speeded-up robust features (SURF)

4.2.3 Visual words integration using BoF model

4.3 Image classification framework

4.4 Experiments and Results

4.4.1 Weighted average of SIFT and SURF

4.4.2 Performance using Corel-1000 image benchmark

4.4.3 Performance on the Corel-1500 image benchmark

4.4.4 Performance on the Corel-2000 image benchmark

4.4.5 Performance on the Oliva and Torralba (OT-Scene) image benchmark

4.4.6 Performance on the GT image benchmark

4.5 Chapter Summary

Chapter 5

5 Conclusion and Future Work

5.1 Conclusion

5.2 Limitations

5.3 Future Directions

References
LIST OF FIGURES

FIGURE 2.1 THE GENERIC FRAMEWORK OF CBIR SYSTEM [1] ................................................................. 6
FIGURE 2.2 METHOD FOR CALCULATION OF CCV [51] ................................................................. 12
FIGURE 2.3 IMAGE ANALYSIS IN TERM OF INTERIOR AND BORDER PIXELS [52] .................. 13
FIGURE 2.4 STATISTICS OF DCD FEATURES FROM 36,692 REGIONS [40] .......................... 14
FIGURE 2.5 TAXONOMY OF TEXTURE DESCRIPTOR [58] ............................................................ 15
FIGURE 2.6 TOP K RETRIEVAL RESULTS [76] .................................................................................. 19
FIGURE 2.7 IMAGE GRADIENT AND KEY POINT DESCRIPTOR [81] ........................................... 20
FIGURE 2.8 DIAGRAM ILLUSTRATING DIFFERENT STEPS OF BOF IMAGE REPRESENTATION MODEL
(FROM FEATURE EXTRACTION TO IMAGE REPRESENTATION AS ORDER-LESS HISTOGRAM OF
VISUAL WORDS [121]) .................................................................................................................... 28
FIGURE 2.9 REPRESENTING THE IDEA OF INVERTED INDEX FOR BOF REPRESENTATION MODEL [121] 30
FIGURE 2.10 TECHNIQUES FOR THE EXTRACTION OF LOCAL FEATURES [127] ............................ 32
FIGURE 2.11 SPATIAL PYRAMID MATCHING (SPM) [126] ............................................................... 36
FIGURE 3.1 IMAGE (A) HAS BEEN CONSTRUCTED BY FOLLOWING THE RULE OF THIRDS AND IMAGE (B)
REPRESENTS A POSSIBLE SPATIAL SEMANTIC SOLUTION FOR EFFICIENT IMAGE RETRIEVAL .................................................... 39
FIGURE 3.2 A DEMONSTRATION OF THE TRIANGULAR RELATIONSHIP BETWEEN OBJECTS AND REGIONS OF INTEREST IN IMAGES FROM THE COREL IMAGE DATASET ............................. 40
FIGURE 3.3 BLOCK DIAGRAM OF THE PROPOSED FRAMEWORK BASED ON HISTOGRAMS OF TRIANGLES [141] .............................................................................................................. 42
FIGURE 3.4 IMAGE (A) REPRESENTS HISTOGRAMS OF TRIANGLES (LEVEL 1) AND IMAGE (B)
REPRESENTS HISTOGRAMS OF TRIANGLES (LEVEL 2) ................................................................... 44
FIGURE 3.5 REPRESENTING THE PROCEDURE FOR THE CALCULATION OF RECTANGULAR HISTOGRAMS
(2X2) ................................................................................................................................................ 48
FIGURE 3.6 SAMPLES OF IMAGES FROM EACH CLASS OF THE COREL A IMAGE BENCHMARK [143] ........ 48
FIGURE 3.7 GRAPHICAL REPRESENTATION OF THE MEAN AVERAGE PRECISION OBTAINED FROM THE COREL A IMAGE DATASET ............................................................................................ 50
FIGURE 3.8 GRAPHICAL REPRESENTATION OF THE MEAN AVERAGE PRECISION OBTAINED FROM COREL A IMAGE BENCHMARK .......................................................................................... 51
FIGURE 3.9 IMAGE RETRIEVAL RESULT OBTAINED BY USING THE PROPOSED IMAGE REPRESENTATION
(LEVEL 2) FOR THE CLASS OF “BUSSES” ..................................................................................... 52
FIGURE 3.10 IMAGE RETRIEVAL RESULT OBTAINED BY USING THE PROPOSED IMAGE REPRESENTATION
(LEVEL 2) FOR THE CLASS OF “BUILDINGS” .................................................................................. 52
FIGURE 3.11 SINGLE LABEL IMAGE ANNOTATION RESULT FOR THE SEMANTIC CLASS “BEACH” .......... 53
FIGURE 3.12 SINGLE LABEL IMAGE ANNOTATION RESULT FOR THE SEMANTIC CLASS “FOOD” ............ 53
FIGURE 3.13 TWO LABELS IMAGE ANNOTATION RESULT FOR THE SEMANTIC CLASS “DINOSAURS” ...... 54
FIGURE 3.14 TWO LABELS IMAGE ANNOTATION RESULT FOR THE SEMANTIC CLASS “FLOWERS” ...... 54
FIGURE 3.15 TWO LABELS IMAGE ANNOTATION RESULT FOR THE SEMANTIC CLASS “HORSES” .......... 55
FIGURE 3.16 THREE LABELS IMAGE ANNOTATION RESULT FOR THE SEMANTIC CLASS “HORSES” ............ 55
FIGURE 3.17 SAMPLES OF IMAGES FROM EACH CLASS OF FIFTEEN SCENE IMAGE DATASET .............. 56
FIGURE 3.18 COMPARISON OF MEAN AVERAGE PRECISION USING FIFTEEN SCENE IMAGE DATASET .. 57
FIGURE 4.1 IMAGES OF DIFFERENT SEMANTIC CLASSES FROM THE COREL IMAGE BENCHMARK ........ 60
FIGURE 4.2 BLOCK DIAGRAM OF VISUAL WORDS INTEGRATION FRAMEWORK OF SIFT AND SURF [156]. ........................................................................................................................................... 63
FIGURE 4.3 SAMPLES OF IMAGES FROM EACH CATEGORY OF THE COREL-1000 IMAGE BENCHMARK [143] .......................................................................................................................................... 69
FIGURE 4.4 COMPARISON OF MEAN AVERAGE PRECISION FOR TOP 20 RETRIEVALS USING THE COREL-1000 ............................................................................................................................ 71
FIGURE 4.5 PRECISION-RECALL CURVE OBTAINED USING THE COREL-1000 IMAGE BENCHMARK ....... 73
FIGURE 4.6 IMAGE RETRIEVAL RESULTS FOR THE CLASS HORSES ..................................................... 73
FIGURE 4.7 IMAGE RETRIEVAL RESULTS FOR THE CLASS DINOSAURS ............................................. 74
FIGURE 4.8 IMAGE RETRIEVAL RESULTS FOR THE CLASS ELEPHANTS ................................................................. 74
FIGURE 4.9 IMAGE RETRIEVAL RESULTS FOR THE CLASS BUSES ................................................................. 75
FIGURE 4.10 SAMPLES OF IMAGES FROM EACH CATEGORY OF THE COREL-1500 IMAGE BENCHMARK [157]. ................................................................................................................................. 75
FIGURE 4.11 COMPARISON OF PRECISION AND RECALL USING THE COREL-1500 IMAGE BENCHMARK 76
FIGURE 4.12 SAMPLES OF IMAGES FROM EACH CATEGORY OF THE COREL-2000 IMAGE BENCHMARK [157]. ................................................................................................................................. 77
FIGURE 4.13 COMPARISON OF PRECISION AND RECALL USING THE COREL-2000 IMAGE BENCHMARK. 78
FIGURE 4.14 SAMPLES OF IMAGES FROM EACH CATEGORY OF THE OT-SCENE IMAGE BENCHMARK [158]. ................................................................................................................................. 78
FIGURE 4.15 COMPARISON OF PRECISION AND RECALL USING THE COREL-2000 IMAGE BENCHMARK. 79
FIGURE 4.16 SAMPLES OF IMAGES FROM 05 CLASSES OF GT IMAGE BENCHMARK USED FOR THE EVALUATION OF THE PROPOSED VISUAL WORD INTEGRATION ..................................................................................... 80
LIST OF TABLES

TABLE 2.1 LIST OF IMAGE BENCHMARKS USED FOR THE EVALUATION OF THE PROPOSED RESEARCH. ........................................................................................................................................26
3.1 AVERAGE RETRIEVAL PRECISION OBTAINED FROM THE PROPOSED IMAGE REPRESENTATIONS BY USING DBN, SVM AND RBF-NN ..........................................................................................................................49
TABLE 3.2 CLASS-WISE COMPARISON OF AVERAGE RECALL OBTAINED FROM COREL A IMAGE DATASET. ........................................................................................................................................50
TABLE 3.3 CLASS-WISE COMPARISON OF AVERAGE RECALL OBTAINED FROM COREL A IMAGE DATASET. ........................................................................................................................................51
TABLE 3.4 COMPARISON OF TIME (IN SECONDS) FOR HISTOGRAM COMPUTATION. ........................................................................................................................................57
TABLE 3.5 TIME REQUIRED FOR RETRIEVING IMAGES. ........................................................................................................................................58
TABLE 4.1 MEAN AVERAGE PRECISION FOR TOP 20 RETRIEVALS (VISUAL WORDS INTEGRATION). ........................................................................................................................................70
TABLE 4.2 MEAN AVERAGE PRECISION FOR TOP 20 RETRIEVALS (WEIGHTED AVERAGE 0.7-0.3). ........................................................................................................................................70
TABLE 4.3 CLASS-WISE COMPARISON OF PRECISION FOR TOP 20 RETRIEVALS. ........................................................................................................................................72
TABLE 4.4 CLASS-WISE COMPARISON OF RECALL FOR TOP 20 RETRIEVALS. ........................................................................................................................................72
TABLE 4.5 COMPARISON OF PRECISION AND RECALL USING THE COREL-1500 IMAGE BENCHMARK. ........................................................................................................................................76
TABLE 4.6 COMPARISON OF PRECISION AND RECALL USING THE COREL-2000 IMAGE BENCHMARK. ........................................................................................................................................78
TABLE 4.7 COMPARISON OF MEAN AVERAGE PRECISION USING OT-SCENE IMAGE BENCHMARK. ........................................................................................................................................80
TABLE 4.8 MEAN AVERAGE PRECISION COMPARISON BY SELECTING 05 CLASSES OF GT IMAGE BENCHMARK. ........................................................................................................................................81
LIST OF ACRONYMS

AIA-Automatic Image Annotation
ANN-Artificial Neural Network
BoF-Bag of Features
BoVWs-Bag of Visual Words
BRISK- Binary Robust Invariant Scalable Keypoints
CBIR-Content–Based Image Retrieval
CCM-Color Co-occurrence Matrix
CDH-Color Difference Histogram
CSD-Color Structure Descriptor
CHKM- Color Histogram for K-Mean
DBN-Deep Belief Networks
DoG-Difference of Gaussian
DCD-Dominant Color Descriptor
DCT-Discrete Cosine Transform
EM-Expectation Maximization
EODH-Edge Oriented Difference Histogram
FD-Fractional Descriptor
FFT-Fast Fourier Transform
GCH-Global Color Histogram
GLCM- Gray Level Co-Occurrence Matrix
GPU-Graphic Processing Unit
HSV- Hue Saturation Value
HOG- Histogram of Oriented Gradients
LSE-Least Square Errors
MLE-Maximum Likelihood Estimation
MRF-Markov Random Field
MSER-Maximally Stable Extremal Regions
MSHP-Most Similar Highest Priority
NN-Neural Networks
QBIC-Query By Image Content
SIFT -Scale Invariant Feature Transform
SURF -Speeded-Up Robust Features
TBIR -Text-Based Image Retrieval
RBF-NN -Radial Basis Function Neural Networks
RBSC- Region-Based vector codebook Sub-band Clustering
RF-Relevance Feedback
RSHD-Rotation & Scale-Invariant Hybrid Image Descriptor
SAR-Simultaneous Auto Regression
SCD-Scalable Color Descriptor
SPM- Spatial Pyramid Matching
SVM-Support Vector Machines
PWT-Pyramid Wavelet Transform
LIST OF PUBLICATIONS


CHAPTER-1

1 Introduction

Due to an increase in the size of storage and a decrease in the price of digital camera, there is an exponential growth in the number of image databases [1-4]. On daily basis, thousands to millions of images are uploaded at different public archives [2]. Digital images play a vital role in a wide range of applications including medical care, education, weather forecasting, social media and art designs [1]. The change of scale, appearance of similar views in the images belonging to the different classes, overlapping objects in the image, changes in illumination and different orientations of the same object make CBIR a challenging research problem [1, 5]. Images can be of a scene containing landscapes, wild life or faces of people. Numerical values in the form of feature vectors are used to represent a two dimensional image and in image classification and retrieval problems, these numerical values are used to predict the class of the image [2].

The content-based analysis of the image provides the useful information to users according to their respective domain. According to the recent literature [1, 2, 5], two approaches are commonly applied to retrieve the images from an image archive that are: Text-Based Image Retrieval (TBIR) [5] and Content-Based Image Retrieval (CBIR) [5]. The first approach is based on the manual text annotation that involves manual labeling of images. This approach provides an option to retrieve the images by using high-level language keywords. The main problem associated with this approach is the difference of human perception as it can label a single image with different keywords [2]. Manual efforts are required to label the images of a larger archive that often lead to misclassifications [1, 6]. These problems make this approach impractical and less-effective for implementation in the real world. In CBIR, visual contents of images are used to retrieve the similar images [5]. Color, texture, shape, spatial layout and local features are commonly used to represent the images in the form of low-level visual features [1, 6]. The comparison of feature vector values determines the output of retrieved images [1].
According to the recent literature [1, 2, 6], CBIR provides an efficient approach to retrieve images as compared to the traditional TBIR [5, 6]. Keeping in view, these facts, we selected CBIR for image search.

1.1 Research Motivation

Since last two decades, there is a rapid increase in the number of multimedia contents. Digital images play a vital role in many application domains. Search for a similar image from an image archive is an open research problem [1]. Humans can describe the images contents in a high-level semantic form while machines can provide less semantic information about images [1]. In CBIR, high-level image visuals like faces, grass, sky etc are represented in the form of low-level feature vectors. The feature vector values contains the visual attributes about the image in the form of numerical values [1]. The representation of an image in the form of low-level numerical values creates a semantic gap between high-level image visuals and low-level feature vector values. The reduction of semantic gap increases the performance of CBIR [2]. The search of images on the internet is based on the manual tags and keywords; human errors and difference in visual perception makes this approach less effective and unreliable [2]. CBIR provides a framework for image search that is based on the comparisons of feature vector values of images and this process do not require human interactions [1].

The image visual attributes are described by color, shape, texture and spatial layout. These attributes are extracted in the form of numerical values. The similarity among these numerical values (feature vectors values) determines the output of retrieved images [1]. The output of CBIR strongly depends on the type of selected image representation/feature space. The discriminating feature vector values can improve the performance of CBIR. Bag of Features (BoF) [7] representation model is commonly used for the image classification and retrieval frameworks. In BoF-based image representation model, an image is represented as an order less histogram of visual words by ignoring the spatial contents of the image. The addition of spatial information to the BoF-based image representation model enhances the performance of CBIR [8, 9].

Color features are used to represent the distribution of color in the images while texture is used to represent a group of similar pixels [5]. The visual feature integration is applied to enhance the
performance of CBIR [1]. A combination of color and texture values integrates the attributes of both features to image representation and it increases the performance of CBIR. The selection of features to be used in a combination is an open research problem as selection of inappropriate features for integration can result in a decrease in the performance of image retrieval [1]. Keeping these facts in view, in this thesis, we address the problems that are associated with image spatial attributes, visual word integration (late fusion) and reduction of semantic gap. The BoF-based classification framework is selected for image retrieval [7]. We evaluated different state-of-the-art classifiers to sort out the best retrieval performance obtained from the proposed image representations.

1.2 Research Objectives

Following are the main research objectives that are related to the main contributions of this thesis:

1. In BoF-based image representation, an image is represented as in order-less histogram of visual words, by ignoring the spatial layout of the image. The spatial layout of the image carries information/details that can enhance the performance of image retrieval. We extracted spatial information from the triangular regions at the time of computation of histograms of visual words. This technique adds the spatial triangular attributes to the inverted index of BoF representation model.

2. High-level image visuals are represented in the form of low-level visual features. The reduction of semantic gap enhances the performance of image retrieval. High-level semantic keywords/ image annotations are associated with images by using triangular histograms to reduce the semantic gaps. One, two and three label image annotations are associated with single image to represent an image in high-level semantic form.

3. In image retrieval, perfect results have not been reported yet because a single feature-based image representation is not robust to all transformations. Among local features, SIFT is more robust to rotation, change of scale, and is capable of capturing local object edges. SURF is reported to be robust to changes in illumination and the SURF descriptor is more distinctive. Due to these fact, we selected the visual words integration of SIFT and SURF to represent an image by using the cluster centers of both descriptors. The
proposed image representation based on visual words integration of SIFT and SURF contains the quantized features of two robust local image representations.

1.3 Research Contributions

This thesis mainly deals with two novel image representations that are mainly based on the image spatial attributes and visual words integration (late fusion of two local features). The first novel image representation adds the spatial attributes that are extracted by computing histograms of triangles by dividing an image into two (Level 1) and four (Level 2) triangles. The division of an image into two and four triangular regions is used for the construction of histograms of visual words. The second novel image representation is based on the visual words integration of Scale Invariant Feature Transform (SIFT) and Speeded-Up Robust Features (SURF). The two robust local features are selected for visual words integration because SIFT is more robust to the change in scale and rotation, while SURF is robust to changes in illumination. The main contributions and objectives of this thesis are:

1. The addition of spatial information to the inverted index of a BoF representation model by computing histograms of visual words from triangular regions within the image.

2. The visual words integration (through late fusion) of Scale Invariant Feature Transform (SIFT) and Speeded-Up Robust Features (SURF).

The detail about addition of spatial information based on histograms of triangular regions is mentioned in chapter 3 while the details about visual word integration of SIFT and SURF is mentioned in chapter 4.

1.4 Structure of the Thesis

This thesis is organized as follows:

**Chapter 2** is about fundamentals of image retrieval and it deals with feature extraction techniques, performance evaluation, and standard image benchmarks and contains detail about
BoF based image representation model and parameters that are used in the research presented in chapter 3 and 4.

**Chapter 3** is about images representation as histograms of triangles and it contains detail about histograms of triangles, experimental results and summary.

**Chapter 4** is about the visual words integration of SIFT and SURF and it contains detail about the visual words integration of SIFT-SURF, experimental results and summary.

**Chapter 5** concludes the thesis. The achievements of chapters 3 and 4 are outlined and it points towards the limitations and future directions of the current research.
CHAPTER 2

2 Fundamentals of Image Retrieval

This chapter deals with the fundamentals of image retrieval and feature extraction. Section 2.1 is about the basic architecture of CBIR, Section 2.2 deals with the common techniques that are used for features extraction, Section 2.3 is about performance evaluation, Section 2.4 is about image benchmarks that are used to evaluate the proposed research, Sections 2.5-2.11 are about Bag of Features (BoF) based image representation model while section 2.12 is about the chapter summary.

Figure 2.1 The generic framework of CBIR system [1].

2.1 Basic Architecture of CBIR

CBIR provides a sustainable solution to retrieve the images that are in a semantic relationship with the query image. Figure 2.1 represents a generic framework of CBIR. The low-level feature vector of a query image is extracted and an image is represented in the feature space. The feature space of a query image is compared with the feature space of the images placed in the
archive/database. The closeness among the feature vector values is calculated by applying any similarity matching that determines the output of retrieved images. The performance of image retrieval relies on the representation of image in the selected feature space and it is also used to determine the output of retrieved images. The detail about feature extraction techniques are mentioned in the following sub-sections.

2.2 Feature Extraction Techniques

An image is a collection of pixels that are arranged in the form of a matrix. In CBIR, images are represented in the form of low-level visual features in the form of numerical values. Feature selection for image representation is an important step as it affects the performance of image retrieval.

2.2.1 Image segmentation

Image segmentation is used for the extraction of homogenous regions from an image. Comprehensive surveys on image segmentation are available in the literature [10, 11]. Different segmentation algorithms are designed by the researchers that are selected according to the requirements of the respective domain. Each segmentation algorithm has its own pros and cons. The common techniques used for image segmentation are: thresholding [12, 13], watershed segmentation [14-18], contour-based [19-21], texture-based segmentation [22, 23], grid-based [24-28], k-means clustering [29-32], statistical model [33, 34], normalized cut [35-37] and region growing segmentation methods [38-41].

Thresholding [12, 13] is a simple type of image segmentation and it can be used to separate the objects from their background. Sezgin et al. [12] presented a comprehensive survey on image thresholding methods. They categorized the 40 thresholding methods into 06 groups that are: object attributes, entropy, spatial correlation, space clustering, histogram shape and local gray-level methods. Kittler method is considered as the best among 40 thresholding methods. It is reported that the separate use of thresholding methods is not reliable [12]. The performance of thresholding methods can be enhanced by using them in a combination and this can be done at feature level by averaging the top values obtained from each method. Due to simplicity of thresholding, it is suitable for the applications that require two-class segmentation problem (need
to separate object from the background). The algorithm performs well on simple images with a bi-modal intensity distribution. However, most of the images do not have bimodal distribution of intensity. In this case, thresholding results cannot partition the images and uneven illumination is another factor that affects the performance of thresholding. Gonzalez et al. [13] proposed adaptive thresholding that can handle this problem by sub-dividing an image into multiple sub-regions. In general, thresholding algorithms do not consider the spatial relationship between pixels and results are quite sensitive to noise. Due to these factors, thresholding is seldom used for image segmentation [2].

Beucher et al. [14] proposed the watershed algorithm and this method is used in the field of mathematical morphology. According to Serra et al. [15], watershed algorithm can be intuitively thought of as topographic relief or landscape that is flooded by water. The height of the landscape at each point represents the intensity of pixels. Watersheds are the dividing lines of the catchment basins of rain falling over the regions. Gradients of the image are input to the watershed transform and highest gradient point represents the catchment basin boundaries [16]. It often results in over-segmentation and is also sensitive to noise. To improve this algorithm, Najman et al. [17] proposed to use morphological operations to reduce over-segmentation. Grau et al. [18] encoded prior information to the algorithm and the cost function is changed from the gradient between two pixels to the difference of posterior probabilities. Moreover, watershed algorithm is poor for detecting the low signal-to-noise ratio [2].

Chan et al. [19] proposed active contour-based segmentation algorithm that can detect objects of an image. The idea of curve evolution and level sets is applied and the proposed technique works efficiently on the objects with no boundary definition. Arbelaez et al. [21] proposed a method for segmentation and contour detection that combined multiple local cues that are dependent on spectral clustering. The output of the contour detector is transformed to the hierarchical region tree by maintaining the quality of contour. The curve stops when it coincides with the object boundary. It has the advantage over the cluster-based segmentation approach as it does not require prior knowledge about the number of clusters [19, 20]. For accurate edge detection, there must be no noise in images and presence of noise can give poor results. This algorithm also requires human interaction for object boundary definition and noise is often a part of digital images. These factors limit the use of contour-based segmentation algorithms in many application domains [2].
Texture-based image segmentation has been proposed by Choe et al. [22] and it is based on Hidden Markov Tree (HMT) and wavelet features. HMT is based on probabilistic graph and it can capture the statistical properties of the wavelet. By using the features of HMT and wavelet, this technique can perform texture classification at different scales. The performance of segmentation is improved by applying Bayesian probabilistic model. Grid-based clustering algorithm is based on the idea of division of an image into sub-blocks [24-28], that are further used to extract the visual features. The main advantage of this technique is the less computational cost. The block-based approach is not useful in the case when a single block is representing two different objects. The block-based approach can be used in domain specific applications such as medical image retrieval [28].

Clustering algorithms are used in cluster-based segmentation, $k$-means is example of popular clustering algorithm [29-32]. In $k$-means algorithm, pixels are divided into different groups of clusters and a region is represented by a group. Usually an image is divided into 4x4 pixel blocks for the extraction of color and texture and same block of cluster represents the similar region. This approach has the drawback as it requires the number of pre-define segments and improper value of $k$ results in a degradation of overall performance. The $k$-means algorithm uses the assumption of data presence in the form of spherical clusters which is not always true [2].

Statistical model-based segmentation is proposed in the literature [33, 34]. Blobworld approach is based on model-based segmentation and feature vectors are used to represent the image. 8 dimensional feature vectors are used for each pixel that represents texture, color and position. Gaussian distribution is applied to model image pixel by using random distributions. Expectation–Maximization (EM) is used to calculate the Gaussian parameters and number of regions for further processing. On the base of parameters, posterior probabilities are used to determine the pixel region relationship which is used for image segmentation.

Comaniciu et al. [35] proposed normalized-cut (Ncut) graph-based segmentation approach. In this approach, an image is represented in the form of a graph. Vertices of the graph represent the pixels of the image and edges weights are used to calculate the similarities of pixels. Idea of disjoint sets is applied that results in minimization of similarity between sets. Each set is considered to be a region and in case of large images, there is increase in graph partition, which is computationally expensive. Tao et al. [36] improved this approach by applying a pre-segmentation and mean shift algorithm. Deng et al. [41] proposed JSEG segmentation that is an
example of region growing algorithm. The color pixels are quantized into classes and it results in creation of a class-map. Region growing technique is applied on the class-map. This technique is more efficient in case of homogenous colors and texture regions.

The selection of segmentation technique is dependent on the type of respective domain. Improper segmentation results can be improved by applying post processing [13, 39] that are helpful in removing the noise.

2.2.2 Color features

Color is one of the fundamental visual feature of digital images and due to advancement in technology, most of digital images contain colors that are located at different levels of pixels. Representation of color is achieved by applying different color spaces and models. Different types of color spaces are proposed in the literature [1, 2], color features are extracted from an image on the basis of respective color model [42]. Some of popular techniques for color features extraction are color moment [25, 43], color histogram [43-46], color correlogram [46-50], Color Coherence Vector (CCV) [51, 52], Scalable Color Descriptor (SCD) [53, 54], Color Structure Descriptor (CSD) [53] and Dominant Color Descriptor (DCD) [39, 40, 55].

Color moments are [25, 43] mainly used for color indexing and similarity matching. They are invariant to rotations and scale changes. Mean, standard deviation and skewness are the examples of first, second and third order color moments, respectively. Vialaya et al. [15] extracted color moments to differentiate between indoor & outdoor images and arranged images in a meaningful hierarchy on the basis of low-level features. Indoor images do not have a uniform color distribution while outdoor images have a uniform color distribution. Color moment is considered as a spatial feature for qualitative attributes representation.

Color histogram and spatial moments are used to extract the global color distribution. Goh et al. [43] improved the image retrieval by organizing the low-level image features in high-level semantic form. Mean shift filtering with uniform kernel is applied for 256x256 gray level images. The output quality is managed by using the bandwidth of selected kernel. Color moments are useful in the case of region-based image retrieval [2]. Color histogram is a simple technique for extraction of color feature and it represents the color distribution of an image [43-46]. Color space is converted into different bins and each color bin has its own
frequency. Color histogram is invariant to translations and rotations. In large image databases, similar color histogram for different objects is possible because in real world there are many objects of same color. Cusano et al. [45] extracted joint histogram in HSV color space and divided an image into sub-blocks by considering the partially overlapping sub-images (as tiles). Huang et al. [46] extracted joint histogram (multidimensional histogram) that are created by using set of local pixel. Goh et al. [43] extracted color histogram by dividing color into 12 bins including the bins for outlier and culture. Color histogram lacks the spatial information about pixels and relationship between different image regions cannot be maintained. Different objects with the same color distribution can results in misclassifications. Color histogram ignores important information like shape and texture and information.

Huang et al. [47] proposed color correlogram for image comparison and indexing. Gray level cooccurrence matrix is the gray scale version of color correlogram. The main feature for color correlogram is the spatial information and it can describe the global distribution of image color and is supervisor to the color histogram [2]. Ojala et al. [49] extracted auto color correlogram (subset of color correlogram) for image retrieval in HSV color space with a visual correspondence that is close to human visual system. Auto color correlogram uses the probability functions to find out the identical colors. Chun et al. [50] proposed CBIR by using a combination of color correlogram and texture features. The size of the color correlogram is minimized by the use of auto correlogram to capture the spatial information between the identical colors.

Color Coherence Vector (CCV) is different from color histogram as it captures the spatial information in the image. It is constructed by improving the performance of Global Color Histogram (GCH). The computational complexity of a CCV is higher than that of color histogram. According to Pass et al. [51], color pixels can be classified as coherent or incoherent. Color region larger than 1% of the image size are considered as coherent and less than 1% are considered as non-coherent. Due to separation of coherent and incoherent pixels and resemblance of significant regions to coherent pixels, CCV performance is better than that of color histogram. Color histogram can be same for two different images while by using CCV difference can be distinguished by separating the histogram of coherent and non coherent region.

Figure 2.2 represents the proposed method for the calculation of CCV [51]. Stehling et al. [52] proposed image retrieval by using interior pixel classification. In the first step of image analysis, quantization of an image in obtained in RGB space and pixels are classified as interior or border
on the basis of the pixel locations by using four neighbors rule. Two histograms are calculated for interior and border pixels. Due to the classification of interior and border pixels, the resulted histograms are discriminative. In-case, there are smaller interior pixels of the same color then there must be some visual property of the image that is useful for creating the difference. Figure 2.3 represents image analysis in term of interior and border pixels. Scalable color descriptor (SCD) belongs to MPEG-7 color descriptor [53] and it uses Haar transform encoding scheme, that is applied on color histogram in HSV color space. Dorairaj et al. [54] proposed an image retrieval by using SCD as color descriptor. Working of a SCD is reported similar as that of color histogram and it is computed in HSV color space by quantization of color space into 256 bins.

![Color Coherence Tree](image)

Figure 2.2 Method for calculation of CCV [51].

Histogram values are truncated into 11 bits integer representation. Higher significance is given to smaller values with higher probability. In-case, if each histogram is to be expressed by using 256 bins and each bin is expressed with 4 bit representation, then it might needs a total of 1024bits/histogram. Haar transform is applied that involves addition and difference operation on adjacent bin values and reduces the size of the descriptor. Low-pass coefficients are obtained by using the sum while high-pass coefficients are obtained by using the difference. High-pass
coefficients are used for a compact representation. After computing the coefficients more subsets are iteratively calculated from the source histogram. Color structure descriptor (CSD) belongs to MPEG-7 standard [53] and is a histogram based descriptor. A structure like element e.g rectangle/square is placed on the image. The performance is dependent on the type and size of moving structure. It requires more computations than a SCD that limits its use in many applications [2].

![Image analysis in term of interior and border pixels](image.png)

The dominant color descriptor (DCD) belongs to MPEG-7 [53] standard and it is used in many research problems [39, 40, 55]. Template based approach is proposed by Liu et al. [38] with the assumption to use the average color of all the pixels. The dominant color is assumed to be the color feature. HSV color histogram is calculated for the regions with quantization into 10, 4 and 4 bins for H, S and V, respectively. The bin with the maximum size is selected and the average values of all pixels of the specific bin are used to define the dominant color. However, it is not practical to assume the number of features in advance and improper segmentation can results in errors. Islam et al. [39] proposed dominant color-based vector quantization algorithm that categorizes the image regions automatically and represents them with a finite set of DCD. The proposed approach is useful as it avoids over-segmentation of regions. According to Zhang et al. [40], the proposed algorithm can be applied to represent an image by 04 dominant colors (shown...
in figure 2.4). Some popular color feature extraction methods are discussed in this section, each one has its own pros and cons and they are implemented in research according to the system requirements.

![Figure 2.4 Statistics of DCD features from 36,692 regions [40].](image)

### 2.2.3 Texture features

Texture is a popular and powerful feature for digital images and is depends on the group of pixels. A comprehensive study on texture features is available in literature [11, 56, 57]. The main requirement for an efficient texture feature is to be invariant to changes in rotation, scale and shape. By extracting texture features, objects with the same shape and color can be distinguishable. It is difficult to deal with texture features as object texture can be irregular, random, rough and smooth. Corner similarities and edges in the images are used to model texture features. Figure 2.5 presents the taxonomy of texture descriptor that is proposed by Tuceryan et al. [58]. Many texture features extraction techniques are proposed in the literature [2], but most of them are not invariant to rotation and shape [2].
These spatial texture features are extracted on the basis of pixels statistics. Tamura features [57], co-occurrence matrix [42, 47, 56], Markov Random Field (MRF) [42, 47, 56], Simultaneous Auto Regression (SAR) [42, 59, 60], Fractional Descriptor (FD) [61] and texton [13, 62] are examples of spatial texture extraction techniques.

![Texture descriptors](image)

**Figure 2.5 Taxonomy of texture descriptor [58].**

Tamura et al. [57] proposed an early approach for the texture classification. There are six perceptual properties of an image that are contrast, coarseness, directionality, linearity, roughness and regularity that are considered as Tamura’s texture features. According to the literature [2], the first three features have gained more popularity while the last three are dependent on the combination of first three. Gray Level Co-occurrence Matrix (GLCM) [42] is an example of statistical texture feature extraction technique and it is used for the gray level images It is used to represent the frequency of any particular pair of gray level in the pair of pixels that are separated at a distance. Statistical properties like entropy, energy, contrast, periodicity and homogeneity are calculated by applying GLCM. The main drawback of GLCM is the high computational cost that limits its use in many application domains [2].

Simultaneous Auto Regression (SAR) is among one of the popular texture model. Long et al. [42] stated that the main reason of popularity of SAR is the less parameters that results in low computations as compared to other methods of MRF. Pixel value is calculated in SAR on the basis of linear combination of the adjusting neighbor pixels. Least Square Errors (LSE) or Maximum Likelihood Estimation (MLE) is applied to calculate the parameters of SAR model. In order to increase the performance of image retrieval, multi-resolution MRF is proposed by Liu et
al. [60] that uses multi resolution Gaussian. The working of MRF is to search for the regular structures in images and MRF is not suitable for any other type of textures measurement in the images. For the working of MRF model, it is necessary to calculate the sufficient neighborhood that is the non-trivial problem. Chaudhuri et al. [61] stated the working of Fractional Descriptor (FD) is on the theory of fractional geometry which is used to describe the pattern and shapes with self similarity. It is used for the calculation of smallest structure that can reproduce the whole shape or pattern. Due to linear relationship between logarithmic box size and logarithmic number of boxes, the measurement of FD is made on the basis of least square fit of two variables. However, roughness can be modeled by using FD and features like contrast and direction are not calculated. FD is not invariant to rotations and it is the main drawback of this method [2].

A set of texture primitives along with their placement rules are used for structural texture-based techniques [13, 62]. Texture elements or texton are the examples of texture primitives. String descriptor is used for textons to measure similarity between two descriptors. According to the recent literature [2], the spatial texture features contain perceptual meaning and they are easy to understand. They have shown good performance in-case of irregular shapes [2]. Most of the spatial texture techniques involve complex searching and optimization process. Noise is the major component of digital images and it can easily affect the performance of spatial texture methods. These methods give poor performance in many applications due to which majority of the work reported in last few years have not considered spatial texture feature [2].

These spectral texture techniques are based on transformation of image to the frequency domain by applying the spatial filter bank. Feature extraction is done at the transformed domain by applying the statistic tools. Due to the use of spatial filter bank, these methods are robust to noise that makes them prominent as compared to spatial texture features. Fast Fourier Transform (FFT) is used for the implementation of these methods that makes spectral methods more efficient than spatial methods. Spectral texture features have the ability to distinguish between varieties of texture features by using their generated features. Most commonly used spectral methods are Fourier transform [63-65], Discrete Cosine Transform (DCT), Wold texture[42, 60], wavelet transform[29, 42, 66-69], Gabor filter [66, 68, 70-72], contourlet transform and curvelet transform [40, 67, 72-76].
Herve et al. [63] applied Fourier Transform (FT) and histograms calculation is made from the FT spectrum with circular and wedged partitions. For image analysis, FT is considered as a powerful tool and it is used for capturing the global image features. According to Ngo et al. [64], low order DCT coefficients are calculated in Mandala space that are used to represent the partial derivative of the referred image. Variance is used for the calculation of texture features and this process is completed by using the derivate images. Lu et al. [65] applied DCT, standard deviation and mean to represent the texture features. Liu et al. [60] and Long et al. [42] extracted the perceptual texture features that are directionality, randomness and periodicity; they created the difference by the use of spectral method in features analysis. They applied inverse Fourier transform for the calculation of auto covariance map of an image. The structure of the image is determined by computing harmonic or random structure by using auto covariance energy ratio. The calculations about harmonic and periodic structure are important as they are used for the extraction of features. In the first case, harmonic structure is used for the calculation of the periodicity of texture and is extracted by applying harmonic peaks from the spectrum of Fourier transform. In the other case, Hough transform is applied that helps in extraction of the directionality by using Fourier transform spectrum. The third texture feature is about the image randomness and is extracted by the use of MRSAR coefficients. This results in ranking of an image with both features and joint rank is determined on the basis of probabilistic weighting. The whole process involves complex types of computations. Keeping these facts in view, whole texture features are not recommended for the development of efficient applications in the domain of image retrieval [2].

According to Liu et al. [38], local features play an important role in image semantic analysis and retrieval. Wavelet transform [29, 42, 66-69] is applied on the images that performs decomposition by using different frequencies components and filters at different levels. Texture features are extracted from each of the decomposed frequency component. In this process, the Tree Structure Wavelet Transform (TWT) and Pyramid Wavelet Transform (PWT) are applied on the images. The main advantage of a wavelet transform over FT is the extraction of texture at local image patches and it works at multiple resolution level. The wavelet transform has the drawback of its sensitivity to the point singularities that is a critical step during the extraction of features. The extraction ability of wavelet is not suitable in case of directional textures [2].
Gabor filter is among popular texture feature extraction technique. Gabor filter ability to capture the texture at multiple directions and scales makes it more beneficial than wavelet transform. Another unique feature of Gabor filter is its optimal localization in both frequency and spatial domain. Dugman et al. [70, 71] predicted the efficient performance of 2D Gabor filter for description of 2D field profiles of cells in cortex of mammals visual system. Manjunath et al. [66] applied Gabor filter in CBIR domain and according to the experimental results, it is better than mostly used feature extraction techniques [2]. Due to robust performance of Gabor filter, MPEG-7 is using Gabor filter as the standard texture descriptor. Zhang et al. [68] proposed the normalization of Gabor filter that makes it rotation invariant among texture features. According to the recent literature [2], the performance of Gabor filter is best among majority of texture feature extraction techniques. It has the drawbacks like it prefers point singularities where edge singularities are required. Arivazhagan et al. [72] reported that the outputs of Gabor filter are not mutually orthogonal and correlation occurs between texture features.

Do et al. [77] proposed contourlet texture descriptor with the uniqueness of its processing that starts from the discrete domain. With parabolic scaling and directional moments, contourlet transformation has achieved optimal approximation rate along the differentiable curves. Moayedi et al. [78] proposed an approach for automatic mass classification of mammograms and divided the method into 03 stages. Contourlet transform is used in the second step for the extraction of contourlet coefficients by the use of genetic algorithm that improved the classification performance of SVM. Limitations of wavelet is addressed by a new approach that is known as curvelet [67, 73], that consists of special combination of wavelet. By applying curvelet transform, it is possible to capture the edges at different scales and orientation. It has shown good results in various domains that are image enhancement, de-noising and image classification [72, 74, 79]. Semeler et al. [80] applied curvelet transformation in the medical domain for the classification. Sumana et al. [75] investigated the second generation of curvelet feature and investigated its performance on small image benchmarks. Zhang et al. [40] selected curvelet as features extraction technique and combined the feature space with DCD and shape descriptor. Islam et al. [76] introduced curvelet transform for image retrieval and proposed rotation invariant features of curvelet transform. According to the experimental results, the performance of curvelet is much better than Gabor filter. The comparison graph obtained from the experimental results [76] is shown in figure 2.6.
The selection of particular feature for texture analysis is dependent on the requirement of respective domain. Each feature has its own pros and cons while spatial features are useful in a way that they can be extracted from an image without losing information they are semantically meaningful. The main drawback of the spatial features is high computational cost and their sensitivity to the noise. According to the literature [2], the spectral features are more robust and require less computational cost [66]. They are implemented by applying FFT, while the main drawback of spectral feature is their non-semantic semantic meaning. Spectral techniques work efficiently on the image with large size while special techniques are useful for the images with smaller size and irregular shapes [2].

### 2.2.4 Interest point detectors

Interest point detectors have shown good results in content-based image matching applications [1]. Scale Invariant Feature Transform (SIFT) [81] and Speeded-Up Robust Features (SURF) [82] and the examples of two robust local features.
Scale space extrema detection, keypoints localization, orientation assignment and keypoint descriptor are the four major steps for computing the SIFT descriptor [81]. In the first step, the Difference-of-Gaussian (DoG) is applied for the calculation of potential interest points and several Gaussian blurred images are produced by applying different scales to the input image. The DoG is calculated by using the neighborhood blur images. A series of DoG is applied to the scale space and stable keypoints are detected by using the maxima and minima of the Laplacian of Gaussian.

In the second step, the extrema are calculated in DoG images for the selection of candidate keypoints. Taylor series is applied to eliminate low contrast and poor localized candidates along the edges. In the third step, the principal orientation is assigned to the keypoints and achieves invariance to image rotation. The fourth step computes the SIFT descriptor across each keypoint. The descriptor gradient orientations and coordinates are rotated relative to the keypoint orientation and provide the orientation invariance. For each keypoint, a set of orientation histograms are created on 4x4 pixel neighborhood, with 8 orientation bins in each. This results in feature vectors containing 128 dimensions, SIFT descriptors are invariant to contrast, scale and rotation [81].

There are two main steps to compute the SURF keypoints and descriptors [82]. The box filter is applied to the integral images for an efficient computation of the Laplacian of Gaussian. Determinants of the Hessian matrix are calculated for the detection of the keypoints. In the second step, every keypoint is assigned to a reproducible orientation by applying the Haar wavelet in the direction of x and y. A square window is applied around the keypoints and is
oriented along the orientations detected before. The Haar wavelets with a size of 2 sigma are calculated by applying the window that is divided into 4x4 regular sub-regions and each sub-region contributes values. This result in feature vectors containing 64 dimensions, SURF descriptors are invariant to rotation, change of scale and contrast [82].

2.2.5 Shape features

Shape plays an important role in feature extraction and comprehensive surveys on shape are available in the literature [83, 84]. According to the requirement of present era, unique identity, affine invariance, noise, tolerance, scale, translation, rotation, less computation power, tolerance for occlusion, efficient retrieval ability from database and reliability are the main requirements while extracting the shape feature. Geometric features are useful as difference or similarities can
be calculated on the basis of geometric calculations. The classification of shape feature is presented in figure 2.8. There popular shape descriptors proposed in the literature are: Hausdroff distance [83, 85], shape signature [86, 87], elastic matching [88], stochastic method [88, 89], scale space method [90-92], Fourier descriptor (FD) [93, 94], Wavelet Descriptor (WD), geometric moment invariant [94-96], algebraic moment invariant [97, 98] and Generic Fourier Descriptor (GFD) [83, 99].

### 2.2.6 Combination of visual features

According to literature, IBM launched first commercial system to retrieved images [1, 3]. Later on, different types of features extraction methods are proposed, that are based on visual features such as color, texture and shape [1, 3, 100, 101]. Local features such as Scale Invariant Feature Transform (SIFT) [81], Histogram of Oriented Gradients (HOG) [102], Speeded-Up Robust Features (SURF) [82] and Maximally Stable Extremal Regions (MSER) [103] are applied in different domains of image retrieval. Lin et al. [104] proposed CBIR by applying a combination of color and texture. Color Histogram for K-Mean (CHKM) is applied to extract color while texture is extracted by the use of the Color Co-occurrence Matrix (CCM). The probability of the occurrence of the same pixel color and its adjacent one is calculated by the use of conventional CCM and is considered an attribute for that image. When color histograms are used, two different images with a similar color distribution result in a degradation of the image retrieval performance [3]. Lai et al. [105] proposed user-oriented image retrieval based on Interactive Genetic Algorithms (IGA). Color is represented by mean value, standard deviation and an image bitmap, while texture is extracted by applying the Gray Level Co-Occurrence Matrix (GLCM) and an edge histogram.

Spectral texture techniques like curvelet, wavelet packets and Gabor features are used in CBIR to reduce the semantic gap [106-108]. Youssef et al. [107] proposed an Integrated Curvelet-Based Image Retrieval Scheme (ICTEDCT-CBIR). A combination of curvelet transform at different orientations is applied with Region-Based vector codebook Subband Clustering (RBSC) for the extraction of dominant colors and texture. The curvelet transform is applied as it restores sparsity by reducing the repeatability across scales. The Most Similar Highest Priority (MSHP) principle is used to search images from the database. Irtaza et al. [106] proposed semantic image retrieval by using wavelet packets with Eigen values of Gabor features. To enhance semantic image
retrieval, back propagation neural network architecture is trained on the output of Relevance Feedback (RF). A combination of wavelet packets, Gabor features and curvelet transform is used [108] for image representation. Yuan et al. [109] proposed image retrieval by applying a combination of Local Binary Pattern (LBP) and SIFT. The approach of applying a combination of SIFT and LBP aims to achieve high performance in the case of noisy background. Visual features are separately extracted from an image by using SIFT and LBP. Wang et al. [9] proposed Spatial Weighting BoF (SWBoF), which uses local entropy, local variance and adjacent block distance. Tian et al. [110] combined color SIFT with Edge Oriented Difference Histogram (EODH) that is robust to changes in scale and rotations. A weighted word distribution is applied to obtain represent the image in the form of two descriptors. Ashraf et al. [111] combined bandlet transform with color features and extracted the objects from the images. An Artificial Neural Network (ANN) is used as a classifier and the inverted index is applied to retrieve similar images.

Karakasis et al. [112] proposed an image retrieval framework by using affine moment invariants as descriptors. The affine moment invariants are extracted with the help of the SURF detector. Wan et al. [113] reported some encouraging results, introducing a deep learning framework for CBIR by training large-scale Convolutional Neural Networks (CNN). According to their conclusions, the features extracted by using a pre-trained CNN model may or may not be better than the traditional hand-crafted features. By applying proper feature refining schemes, the deep learning feature representations consistently outperform conventional hand-crafted features.

Lenc et al. [114] combined the descriptors of SIFT and SURF for Automatic Face Recognition (AFR). The framework is based on early features fusion of SIFT and SURF. According to Liu et al. [115], spatial information carries significant information for content verification. The spatial context of local features is represented in binary codes for implicit geometric verification. According to the experimental results, the multimode property of local features improves the efficiency of image retrieval. Guo et al. [116] proposed Dot-Diffused Block Truncation Coding (DDBTC), which is based on a compressed data stream, in order to derive image feature descriptors. A DDBTC-based color quantizer and its correspondence bitmap are used to construct the feature space. An image compressed by applying DDBTC provides an efficient image retrieval and classification framework. Liu et al. [117] organized the local features into dozens of groups by applying k-means clustering. In this approach, a compact descriptor is
selected to describe the visual information of each group. This reorganization of thousands of local features into dozens of groups reduces complexity for a large-scale image search. Yildizer et al. [118] proposed CBIR for non-texture images and applied Daubechies wavelet transformation to divide an image into high and low frequency bands. The multi-class Support Vector Regression (SVR) model is applied to represent the images in the form of low-level features. To improve the performance of image retrieval, Yu et al. [119] integrated SIFT and HOG with LBP. The visual features of SIFT and LBP are extracted separately. To obtain a hybrid image representation, the weighted average $k$-means clustering is applied to maintain a balance between the extracted feature space. According to the experimental results, the best retrieval performance is obtained by using the features integration of SIFT and LBP [119].

### 2.3 Performance Evaluation

According to the recent literature [1], different performance measures are applied to evaluate the research of image retrieval. Selecting a particular measure for the evaluation of image retrieval is dependent upon the associated problem domain. The detail about the set of common measures used in evaluation of image retrieval research is discussed in the following sub-sections.

#### 2.3.1 Precision and recall

Precision and recall are the common measures that are used to evaluate the performance of CBIR. Precision determines the number of correctly retrieved images.

$$Precision = P_k = \frac{Kr}{Nr}$$ \hspace{1cm} (2.1)

where $Kr$ represents the number of relevant images similar to the query and $Xr$ indicates the number of images retrieved by the system in response to the query.

Recall is the ratio of correct images retrieved to the total number of images of that class in the dataset.

$$Recall = R_k = \frac{Xc}{Nr}$$ \hspace{1cm} (2.2)
where Xc is total number of images of that class in the database.

2.3.2 F-measure

The harmonic mean of precision and recall is known as F-measure or F1-score and it is calculated by using the following equation:

\[
F1 = 2 \times \frac{P_k \cdot R_k}{P_k + R_k} \tag{2.3}
\]

2.3.3 Average-precision

Average-precision (AP) is a global measure to evaluate the performance of CBIR. For a query q, the AP is the mean of precision value for each relevant image.

\[
AP_q = \frac{1}{Nq} \sum_{k=1}^{N_{VR}} P_k (R_k) \tag{2.4}
\]

2.3.4 Mean average precision

The Mean Average Precision (MAP) is the mean precision that is calculated over all queries and is represented as:

\[
MAP_q = \frac{1}{Nq} \sum_{k=1}^{N_{VR}} AP_q \tag{2.5}
\]
2.3.5 Precision-recall curve

Precision–recall curve is used to plot and compare the set of precision and recall values by using a graphical curve. This provides an option to evaluate the complete retrieval performance in the form of curve that is constructed by using the values of precision and recall.

<table>
<thead>
<tr>
<th>Name of Dataset</th>
<th>Total classes</th>
<th>Total images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corel-1000</td>
<td>10</td>
<td>1000</td>
</tr>
<tr>
<td>Corel-1500</td>
<td>15</td>
<td>1500</td>
</tr>
<tr>
<td>Corel-2000</td>
<td>20</td>
<td>2000</td>
</tr>
<tr>
<td>Oliva and Torralba (OT-Scene)</td>
<td>08</td>
<td>2688</td>
</tr>
<tr>
<td>Scene-Fifteen</td>
<td>15</td>
<td>4485</td>
</tr>
<tr>
<td>Ground Truth (GT)</td>
<td>22</td>
<td>1109</td>
</tr>
</tbody>
</table>

Table 2.1 List of image benchmarks used for the evaluation of the proposed research.

2.4 Image Benchmarks/ Datasets

The image benchmarks/datasets selected for the evaluation of the proposed research are: Corel-1000, Corel-1500, Corel-2000, Oliva and Torralba (OT-Scene), Scene-Fifteen and Ground Truth (GT). The detail about the total number of classes and images in the selected dataset used for the evaluation of the proposed research in mentioned in the table 2.1. The reason behind selection of these image dataset is their widely used in the research that is selected for a purpose of performance evaluation.

2.5 Bag of Features (BoF) Model

The proposed research is based on BoF-based image representation model with image classification framework. In BoF-based image representation model, an image is named with a specific category that represents the class of that image. Image of a given class for-example car
will be different from another class of transportation with the common features like doors and wheels. The common features are usually selected to perform learning of any classification-based model. With the help of a trained model, class labels are associated with the images that contain instances of the same object. The same is the case with text-based analysis where text labels are required to assign to a topic like sports-event or international-news. Codebook learning is the solution that is being proposed to handle such cases [7, 120]. The training dataset is used for the learning of the classifier that consists of words that are related to a particular topic. After learning, a classifier is able to assign any label to the unseen data.

Images consist of different patches and a group of patches from different viewpoints, change of scale and rotation are used to predict that the class of a query image is related to any particular patch. A visual vocabulary (also known as codebook) consisting of image patches can be constructed in a similar way as the text vocabulary (codebook) that is used for text analysis [121]. This approach is known as Bag of Features (BoF) or Bag of Visual Words (BoVWs) [7]. Figure 3.1 represents the different steps from features extraction to image representation as an order-less histogram of visual words.

According to the literature, Csurka et al. [122] proposed the BoF-based image representation with image classification-based framework. Later on, it is being modified in different research applications and evaluated by using different image benchmarks [1]. In BoF-based image representation model, the feature vectors of the images are the respective histograms of visual words. A classification algorithm like SVM [123] or Neural Network (NN) [124] can be used to perform image classification. Training dataset is selected to train the classifier and performance of classifier is evaluated on the test dataset with unseen data. In case of supervised learning, different labels are associated with training images and each label is associated with particular class. Based on the training images/ training data, a classifier predicts the labels of the data that belongs to the test dataset.

The classifier assigns class label to each image and the classifier scores/probability values are used to predict the closeness among the images with in the same class. Image retrieval is performed by calculating the distance of the query image to the images placed in an archive. Local features like SIFT [81] or SURF [82] are usually selected to extract the image patches. Following are the main steps that are associated with image retrieval based on the BoF representation model.
Figure 2.8 Diagram illustrating different steps of BoF image representation model (from feature extraction to image representation as order-less histogram of visual words [121].

### 2.5.1 Extraction of local features

Local features (like SIFT, SURF etc) are extracted from a set of images and this process is known as features sampling [122, 125]. Local features can be extracted by using a dense grid that extracts features by using regular specified steps. Local features can also be extracted by using interest point-based detectors [81]. Another sampling technique is the combination of dense and interest point detector and is known as dense interest point sampling.
2.5.2 Clustering

The second step is associated with the clustering of local features. The high dimensional feature space is clustered by applying a quantization algorithm like $k$-means to construct the visual vocabulary that is also known as code-book.

2.5.3 Encoding of feature space

Local features are extracted from a given set of images and then quantized; visual words are assigned to the image by calculating the Euclidean distance between the visual words and the quantized descriptor. An image is represented as a histogram of words and each bin of the histogram corresponds to the respective visual word of that image. The size of the histogram represents the size of the visual vocabulary.

2.5.4 Inverted index for BoF model

In BoF-based image representation model, an inverted file indexing is used to store the images. Inverted indexing is similar as the index of a book and keywords are used to map the page numbers. In BoF model, a table is constructed that indicates the occurrence of visual words. The extracted information is added to the inverted index of BoF representation model. Figure 3.2 represents the same idea of inverted index. The image on the left side contains the image indexing while the image on the right side represents the idea of image retrieval on the basis of similarity measures between the respective visual words.

2.6 Image Representation Using BoF Model

In BoF model, an image is represented as an order less histogram of visual words. The order less image representation provides a flexibility to the pose and view point change [7]. The spatial
information provides discriminating details in classification and retrieval problems. Different techniques are proposed in the literature to add the spatial attributes for an efficient image retrieval system [8, 126].

![Image](image.png)

Figure 2.9 Representing the idea of inverted index for BoF representation model [121].

### 2.6.1 Image Classification

The histograms of visual words that are constructed by using BoF based image representation are normalized. A classification algorithm like SVM or ANN is applied on the normalized histograms of visual words to perform classification.

### 2.6.2 Distance measures

The classifier assigns class label to a set of images. The class label predicts the class of an image while the classifier score determines the closeness of images with in the same class. The Euclidean distance between a query image to the images placed in an archive determines the output of retrieved images.
2.7 Types of Local Features

According to the literature [122, 127], the learning model of any classifier is strongly dependant on the type of selected image representation and selection of improper features degrades the performance. In BoF-based image representation, local features are extracted in the first step. The detail about the extraction of local features is divided into following types [125, 127].

2.7.1 Interest point-detectors/Sparse features

There are different regions in an image; the regions with more information contents can be extracted by applying interest point-based sampling. The interest point-based sampling is invariant to illumination, changes in the view point and rotations. Figure 3.3 (the image on the right side) represents an example of interest point-based sampling. It can be seen that the interest point detector has localized the interesting regions [127] that are located in the image of the person and car. According to the literature, the interest point detectors performance is satisfactory incase of matching applications and they are not reported efficient with the problems that are based on classification framework [125, 128, 129]. The number of interest points varies according to the image contents and incase of low contrast images, interest point detectors fail to find a single point that make its use impractical for the problem that are based on classification frameworks.

2.7.2 Dense features extraction

Dense feature extraction/sampling extracts the features by using a specified grid [126, 128, 130]. Local features are extracted by applying a regular grid with a fixed step size/pixel stride. Pixels are arranged in an image in the form of row and columns, for a dense pixel stride of 5, every 5th pixel is considered as a feature to compute the descriptor. For-example figure 3.3 (the image on the left side) represents an example of dense features extraction. The main benefit of dense sampling as compared to the interest point-based detector is that the equal importance is given to the low contrast regions. Dense features can also capture the weak spatial information among the
local features. Dense sampling is computational expensive and the overlapping of regions among local features do not guarantee similar feature descriptor [125, 127].

Figure 2.10 Techniques for the extraction of local features [127].

2.7.3 Random features extraction

According to the experimental results [125], random sampling is applied for the extraction of local features on multi-scale and it performs better than interest point-based sampling. The results obtained from dense sampling are much better than that of random sampling [125].

2.7.4 Dense interest point sampling

This is a combination of dense and interest point-based sampling [127]. In this case dense features are extracted on a grid by applying multiple scales. The pixel step size (also known as pixel stride) is arranged in a way to minimize the overlap of image patches. For-example figure
3.3 (the image in the center) represent an example of dense interest point sampling. The main focus of dense interest point sampling is around the interesting regions of the image like interest point sampling.

2.8 Clustering

Local features are clustered to construct a visual vocabulary (also known as codebook). The clustering is beneficial as it reduces the high dimensions of feature space. High computational cost is required to perform clustering and system performance depends on the pre-defined number of clusters. The clustering technique is mainly divided into two groups.

2.8.1 Hard clustering

In hard clustering, local features are assigned to a single cluster by calculating the similarity measure like Euclidean distance. K-means is an example of hard clustering and is commonly used in BoF-based image representation model [7, 131]. The feature space is quantized into informative regions and a region is represented by calculating means of all points. The means are known as visual words and a collection of visual words represents a visual vocabulary.

2.8.2 Soft clustering

In soft clustering, local features are represented by clusters by using a weighted manager. More than one cluster can be assigned to a descriptor by calculating the probability distribution. A Gaussian Mixture Model (GMM) [132] is applied to construct such soft vocabulary. Expectation Maximization (EM) [133] is used to learn the parameters for GMM model.

2.9 Encoding of Feature Space

Local features are extracted from a set of images and quantized in the feature space. Visual words from the vocabulary are assigned to the extracted local features by calculating the Euclidean distance between the visual words and the quantized descriptor [7, 126]. The image is represented as a histogram of visual words and each histogram counts the instances of respective
visual word. The size of the histogram represents the size of vocabulary. According to recent research, Fisher vector encoding and soft quantization [131] represents an image in the form of histogram of visual words in more sophisticated manner. A brief summary about Fisher vector encoding and soft quantization is mentioned in the following sub-sections.

2.9.1 Fisher encoding

In fisher encoding [134], a soft vocabulary is constructed by applying a GMM and descriptors are represented as vectors. The entries of vectors are the first and second order difference between the descriptor and center of a GMM.

2.9.2 Soft quantization

The local features are identical for the images with close visual appearance. Rotation, changes in illumination and noise can result in variation of the respective descriptor. Hard clustering technique like k-means applied to such images can result in assignment of same descriptor to two different clusters. Soft quantization is used to handle such cases and it can assign multiple visual words to a single feature and a weighted scheme is applied that is based on the distance of cluster center and given descriptor.

2.10 Addition of the Spatial Information

The main problem using the BoF-based image representation is the lack of spatial information as images are represented in an order-less histogram of visual words by ignoring the spatial contents [131]. This property of BoF-based representation provides flexibility to the changes in the pose and viewpoints. According to the literature [126], the spatial information among visual words provides discriminating details in retrieval problems. Two techniques are commonly applied to add the image spatial attributes to the inverted index of BoF representation model. The first technique is based on the visual word co-occurrences [135, 136] while second technique is to split and image into different regions and construct histogram of visual words from each of the divided region [126].
2.10.1 Co-occurrence of visual words

In visual words co-occurrence, the spatial information is captured by using the positions of visual words [136]. According to literature, Savarese et al. [135] applied the correlogram of visual words to capture the spatial attributes of images. Correlogram consists of three dimensions that contain the number of times two visual words occur together at a particular distance with respect to each other. The histograms of correlations among visual words are used as an input for classification. Different types of visual word co-occurrences are proposed to enhance the performance of content-based image matching applications. The main drawback of this approach is the high computational cost with the vocabulary of large size [137].

2.10.2 Division of image

This technique divides an image into sub-regions and constructs histograms of visual words from each of the divided region and adds spatial information to the inverted index of BoF representation model. The most popular work associated with the constructing of histogram of visual words from image sub-regions is known as Spatial Pyramid Matching (SPM) [126]. Figure 3.2 represents three levels of SPM that is based on spatial pyramid match kernel. There is no division in case of level0, while in level1 an image is divided into four regions and for level2 image is divided into sixteen regions. Histograms of visual words are constructed from each of the divided regions of grid by applying the weighting scheme as mentioned in the figure 3.2. Keeping in view, the robust performance of this technique, different schemes for image representation are proposed that are associated with construction of histograms of visual words from different regions of image.
Figure 2.11 Spatial pyramid matching (SPM) [126].

2.11 Image Classification

The histograms of visual words that are constructed by using BoF-based image representation are normalized and a classification algorithm like SVM is applied to perform the classification. A group of training images are required to train the classifier. After learning, a classifier is able to predict the class of unseen images. Classification is divided into three main groups that are classified according to the features they employ [138]. The first group is associated with the appearance-based methods that use the features that are relevant to the object color, texture and shape. The second group is associated with the geometry such as edges and lines. The third group is based on local invariant features methods like SIFT, following are the main advantages of using local features [138] with classification-based framework.

- Object models can be learnt without user intervention, user is required to provide the exemplar images that contain the object.
- No geometric primitives are required.
- Can handle occlusion and work efficiently on objects with complex background.
The classifier assigns class labels to a set of images. The class label predicts the class of an image while the classifier score determines the closeness of images with in the same class. The Euclidean distance of a query image to the images placed in an archive determines the output of retrieved images.

### 2.12 Chapter Summary

Keeping in view, the image representation in the form of an order-less histogram we selected the triangular regions (chapter 3) for the extraction of histograms. This procedure adds the spatial information to the inverted index of BoF representation model. Dense features sampling are reported efficient for natural scene classification-based problems. We extracted dense features from triangular region of images and $k$-means is used for clustering. For classification-based image retrieval framework, the classifier performance affects the retrieval precision so we evaluated different classifiers to sort out the best retrieval performance. The classifiers output label is used to predict the class of the images while the similarity of probability scores among the images belonging to the same class determines the output of retrieved images.

According to recent literature [1], the performance of local features are reported robust as they represent image in high-dimensional feature space. Among local features, SIFT is more robust to rotation, change of scale, and is capable of capturing local object edges and shape by using the distribution of the intensity gradients. SIFT performs accurately on the images with a simple background and represents them without noise interference. The performance of SIFT decreases with a complex noisy background and changes in illumination. SURF is reported to be robust to changes in illumination and the SURF descriptor is more distinctive. Due to the robust performance of two local features (SIFT & SURF), we evaluated the visual words integration of SIFT and SURF for an efficient and effective image retrieval. In the second problem (presented in chapter 4), we applied interest point detectors to develop a system that is robust to many transformations. Due to efficient performance we selected SIFT and SURF for feature extraction. We selected a non-linear SVM kernel (Hellinger) for classification to map the data in a high-dimensional feature space. The closeness among the classifier scores with in the same class are used to retrieve similar images.
Chapter- 3

3 Histogram of Triangular Regions

The compositional and content attributes of images carry information that enhances the performance of image retrieval. Standard images are constructed by following the rule of thirds that divides an image into nine equal parts by placing objects or regions of interest at the intersecting lines of the grid. An image represents regions and objects that are in a spatial semantic relationship with respect to each other. While the Bag of Features (BoF) representation is commonly used for image retrieval, it lacks spatial information. In this chapter, we present two novel image representation methods based on the histograms of triangles, which add spatial information to the inverted index of BoF representation. Histograms of triangles are computed at two levels, by dividing an image into two and four triangles that are evaluated separately. Extensive experiments and comparisons conducted on two datasets demonstrate that the proposed image representations enhance the performance of image retrieval.

3.1 Introduction

Content-Based Image Retrieval (CBIR) provides a sustainable way to search for similar images in image archives by using the visual features of a query image [1, 3]. Rotations, change of scale, change of viewpoint, difference is image resolution, overlapping objects in images and the exponential increase in image databases all contribute to make CBIR an active research area. [1, 3]. In CBIR, feature vectors are used to represent images in the form of low-level visual features [1, 3]. The feature vector of a query image is calculated and compared with the feature vectors of the images placed in an archive [106]. The difference of feature vector values of a query image to the images placed in an archive determines the output of retrieved images. Images belonging to the class Beach or Mountains can have same visual appearances like sky or clouds. As a result, there is a closeness among the values of respective image feature vectors and it result in the output of irrelevant images [106]. In an image, there is a spatial relationship between different regions of the image, yet due to the order-less representation of the image
using BoF [7], this spatial information is lost. Approaches based on query expansion[139], soft quantization [131] and large vocabulary size [124] are used to enhance the performance of image retrieval. The main limitation of all these approaches is the lack of spatial information that provides discriminating details. Two approaches are commonly used to add spatial information to the BoF representation [137]. The first approach uses geometric relationships [110] or the co-occurrence of visual words [9]. However, modeling these approaches in the case of large vocabulary size is computationally expensive [137]. The second approach [109] splits an image into sub-regions and constructs the histograms from each sub-region; it is considered robust for content-based image matching [109]. In view of the robust performance of the second approach, our proposed work involves extracting the histograms from triangular regions of the image.

Figure 3.1 Image (a) has been constructed by following the rule of thirds and image (b) represents a possible spatial semantic solution for efficient image retrieval.

Images are constructed by following the rule of thirds, which divides an image into nine equal grids; objects or regions of interest are placed at intersecting lines of the grid (either on the left or on the right). The rule of thirds represents compositional and content attributes of the image, and incorporating these attributes into image retrieval enhances its performance [140]. Figure 3.1 (a) represents a standard image constructed by following the rule of thirds, while Figure. 3.1 (b) represents a possible spatial semantic solution for efficient image retrieval. In a standard image, the sky, the sun or clouds are located at the top, objects of interest are at the right or left, and the
ground, grass or water are found at the bottom of the image. This sequence represents a triangular relationship between objects and regions of interest in a scene. Figure 3.2 represents this content-based triangular relationship between different images of the Corel dataset. Discriminating objects (such as horses or elephants) and regions of interest (such as the sky, the ground or grass) are usually located in different sub-regions of the image. The construction of histograms from triangular regions of the image adds discriminating information to image retrieval in the form of objects and regions of interest that are located at the top, left, right and bottom of the image.

Figure 3.2 A demonstration of the triangular relationship between objects and regions of interest in images from the Corel image dataset.

Inspired by this, we propose two novel methods of image representation in the form of histograms computed from the triangular regions of an image. Dense features are extracted from an image; then, the feature space is quantized and visual vocabulary is constructed in order to achieve a compact representation. An image is divided into two and four triangular regions, while histograms are computed from the triangular regions. Images are divided into two and four triangular regions, which are referred to as Level 1 and Level 2, respectively. Here are the main contributions of this chapter:
1. The addition of spatial information to the inverted index of the BoF representation.

2. Image representation in the form of triangular histograms.

### 3.2 Bag of Features Based Image Representation

In the BoF representation [7], local features are extracted from a set of training images. These features contain information about images patches. The extracted feature space is large in number, in order to reduce the space; a quantization algorithm (such as k-means [7]) is applied on training image features. As a result, cluster centers (also known as visual words) are created and the combination of clusters/visual words are used to represent the constructed visual vocabulary. In order to represent the images using constructed clusters/visual words, the local features are extracted from the image and feature space is quantized, visual words are assigned to the image by using the Euclidean distance between the visual words and the quantized descriptor. The co-occurrences of visual words are counted to represent the images in the form of order-less histogram by ignoring the spatial contents. The dimensions of represented histograms are equal to size of quantized feature space. Due to order-less image representation using BoF representation there is a flexibility in viewpoint changes [1, 7].

#### 3.2.1 Computation of triangular histograms

The order less representation of the image is a drawback in BoF representation model [1, 124, 137]. However, spatial information provides discriminating details in classification based problems [8, 124]. The block diagram of the proposed triangular approach is presented in Figure 3.3.

1. In BoF model, an image \( R \) is represented as:

\[
R = (a_{m,n})
\]  \hspace{1cm} (3.1)

Where \( m \) and \( n \) are the co-ordinates along \( x \) and \( y \) axis respectively.
Figure 3.3 Block diagram of the proposed framework based on histograms of triangles [141].
2. A regular grid is applied to extract the dense features from the image and are represented as:

\[ R = \{d_1, d_2, d_3, \ldots, d_n\} \]  

(3.2)

Where \(d_1\) to \(d_n\) are the descriptors of image \(R\).

3. A clustering algorithm such as \(k\)-means is applied to reduce/quantize the feature space into \(t\) number of words that are represented as:

\[ \text{voc} = \{v_1, v_2, v_3, \ldots, v_t\} \]  

(3.3)

Where \(v_1\) to \(v_t\) are the visual words (or clusters) while \(\text{voc}\) represents the constructed visual vocabulary (codebook with \(t\) clusters).

4. To represent an image in the form of visual words by using the constructed visual vocabulary, dense features are extracted and quantized in the feature space. Triangular regions are selected for the mapping of visual words. An image is divided into two and four triangular regions for Level 1 and Level 2 triangles. The following equation is used to assign the visual words to the reduced/quantized feature space.

\[
v(d_k) = \arg\min_{v \in \text{voc}} \text{Dist}(v, d_k)
\]  

(3.4)

Where \(v(d_k)\) is the visual word that is assigned to the \(k\)th descriptor \(d_k\) and the Euclidean distance between the feature space and visual word is represented by \(\text{Dist}(v, d_k)\).

5. The histograms of \(N\) visual words (equal to the size of constructed visual vocabulary) are extracted from each of the triangular regions of an image (Level 1 and Level 2). Two histograms of \(N\) visual words are extracted from Level 1 and four histograms of \(N\) visual words are extracted from Level 2. The histograms of triangles (Level 1 and Level 2) add spatial information to the inverted index of BoF representation. The representations of the image as histograms of the Level 1 and Level 2 triangles are shown in Figure 3.4.

6. The image representation as histogram of triangles is mathematically expressed as:
\[ H(L_j) = \sum_{j=1}^{n} \frac{R}{j} \quad (3.5) \]

where \( H \) is the histogram of image \( R \), and \( L_j \) represents the level of divided triangles.

7. The detail about dense features extraction and vocabulary construction is mentioned in section 3.4.

Figure 3.4 Image (a) represents histograms of triangles (Level 1) and image (b) represents histograms of triangles (Level 2).

### 3.3 Image Classification Framework

The histograms constructed from the triangular regions of the images are normalized, and classification is performed by using Support Vector Machines (SVM) \([123]\) (with Hellinger kernel \([108]\)), Radial Basis Function Neural Networks (RBF-NN) \([124]\) and Deep Belief Networks (DBN) \([142]\), which consist of autoencoders. Three different classifiers are selected in order to evaluate the best performance of the proposed image representations.
3.3.1 Support vector machines

Support Vector Machine (SVM) is a state of art supervised learning classification method [2]. The kernel method [2] enables SVM to compute the dot product in a separable feature space that helps to generate a decision boundary for non-linear function. The histograms of visual words constructed by using level 1 and level 2 triangles are normalized by applying L2 normalization (to use the relative contribution of histograms bins regardless to the absolute values). The SVM Hellinger kernel [139] is applied on the triangular histograms by applying following equation:

\[ K(z, z') = \sum_i \sqrt{z(i)z'(i)} \]  

(3.6)

Where \( z(i) \) and \( z'(i) \) are the triangular histograms of visual words.

We applied the Hellinger kernel function as its works on the computational cost of linear SVM. The one-against-one rule is applied and for \( t \) number of classes, \( t \cdot (t-1)/2 \) classifiers are constructed. The maximum score value of an image is used to predict the class label. 10 fold cross-validations are applied on training histograms to determine the regularization parameter (\( C \)) of linear SVM.

3.3.2 Radial basis function neural networks

An Artificial Neural Network (ANN) is considered as a powerful tool and is used in pattern recognition and classification based problems [124]. ANN consists of multiple layers of interconnected nodes that learn about multiple classes at a time. The first layer of an ANN is called input layer and the numbers of neurons in the layer are consider equal to the size of visual vocabulary. The numbers of neurons in the output layer are equal to the total number of classes. The histograms of visual words constructed by using level 1 and level 2 triangles are normalized by applying L2 normalization and RBF-NN (by using default RBF-function) is applied on the normalized histograms of visual words.
3.3.3 Deep belief networks

The Deep Belief Networks (DBN) consists of a stack of auto-encoders that are trained by applying unsupervised learning. The stack of auto-encoders is combined to create soft-max layer that is used for the classification [142]. The histograms constructed by using the BoF representation are normalized by applying L2 normalization and DBN consisting of two stacked auto-encoders is applied on the normalized histograms of visual words. The optimization of DBN is performed by varying the number of hidden layers of DBN.

3.4 Experimental Parameters

The detail about the experimental parameters used for the evaluation of the proposed research is mentioned in this section. The proposed image representations are evaluated on two image benchmarks (Corel A and Fifteen Scene). We applied k-means unsupervised clustering to build/construct the codebook therefore, all experiments are repeated 10 times and the average (mean) values are reported. For every run training and test datasets are selected randomly. Classifier decision label is used to determine the class of the image while classification decision score/probability score is used to sort out the similarity/closeness between query image and images placed in the test dataset. 50% of random features from the training dataset are used for the construction of visual vocabulary.

The performance of image retrieved based on BoF model depends on the visual vocabulary size or number of cluster in the codebook. An increase in the vocabulary size/number of clusters in the codebook results an increase in the mean performance while on larger vocabulary size there is a decrease in mean performance due to over-fitting [125]. Visual vocabularies of different sizes are constructed to determine the best performance of the proposed image representations. Randomly selected training images dataset are used to construct the visual vocabulary and results are evaluated by using the test dataset (based on a random selection). All the images are processed in standard size as available in the respective dataset. Training and testing is performed by a random split of 70% and 30% respectively. The images are processed in the standard size as available in the respective dataset.
Dense SIFT with a bin size of 8 (to avoid feature overlapping) and a step size of 10 (to avoid the high computational cost that is associated with vocabulary construction on a smaller step size) [125] is used for the features extraction and all the processing is performed using gray scale images. The step size is also known as pixel stride and it is applied to control the spatial resolution of the dense grid. A smaller step size value results in finer grid. The larger step size makes the dense grid coarser. For a step size of 10, we computed SIFT descriptor after every 10\(^{th}\) pixel. 50% of features from a set of training images are used to construct the visual vocabulary (codebook/number of clusters). Classifier decision labels/probability scores are used to attach semantic labels/image annotations [2] to represent image in the form of keywords. The pre-defined semantic class labels that are used for the 10 classes of Corel A image dataset are: African, Human, Beach, Water, Buildings, Art, Busses, Vehicle, Dinosaurs, Wild-life, Elephants, Forest & tree, Flowers, Petals, Horses, Plants & grass, Mountains, Land area, Food and Eatables. AIA results are reported by associating one, two and three labels to a single image. In other words, a single image is associated with more than one class for the purpose of image annotation [2]. Single and two label annotation is achieved on the basis of top-classification score of the same class. For three level annotations, first two text labels/annotations are predicted on the basis of the classifier assigned class to the image while the third label/annotation is predicted by considering the second top classification score.

Lazebnik et al. [109] proposed Spatial Pyramid Matching (SPM), which divides an image into several rectangular grids and constructs histograms from each region of the grid. The proposed image representations based on triangular histograms are compared with rectangular histograms that are constructed by dividing an image into 2x2 rectangular grids. The representation of an image as rectangular histograms (2x2) is represented in Figure. 3.5. The 2x2 rectangular grid (with four rectangular regions) is selected to perform a comparison with the histograms of Level 2 triangles (with four triangular regions).
3.4.1 Performance on the Corel A image benchmark

Figure 3.5 Representing the procedure for the calculation of rectangular histograms (2x2).

Figure 3.6 Samples of images from each class of the Corel A image benchmark [143].
The Corel A image benchmark [143] is used for the evaluation of the proposed image representations. The results are compared with state-of-the-art research [9, 106, 111]. The Corel A image dataset contains 1000 images (100 images per class) that are divided into ten semantic categories, namely: Africa, Buildings, Beach, Dinosaurs, Buses, Elephants, Horses, Flowers, Mountains and Food. Fig.3.6 shows the images from all of the categories from the Corel A image dataset. In order to sort out the best retrieval performance, we test various sizes of vocabulary [50, 100, 150, 200, 300, 400, 500]. The best mean average precision is obtained from the proposed image representations on a vocabulary with a size of 200. The results are reported for a random selection of 300 images from the test dataset. The average retrieval precision for the top 20 retrievals is calculated by using DBN, SVM and RBF-NN and is shown in Table 3.1. The mean average precision obtained from each of the classifiers is graphically represented in Figure. 3.7.

<table>
<thead>
<tr>
<th>Class/Method</th>
<th>Level 2</th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DBN</td>
<td>DBN</td>
<td>SVM</td>
<td>DBN</td>
<td>ANN</td>
<td>ANN</td>
</tr>
<tr>
<td>Africa</td>
<td>78.34</td>
<td>65.12</td>
<td>69.08</td>
<td>65.12</td>
<td>78.23</td>
<td>73.06</td>
</tr>
<tr>
<td>Beach</td>
<td>82.44</td>
<td>71.09</td>
<td>72.20</td>
<td>71.09</td>
<td>75.92</td>
<td>69.98</td>
</tr>
<tr>
<td>Building</td>
<td>91.62</td>
<td>84.16</td>
<td>84.85</td>
<td>84.16</td>
<td>76.00</td>
<td>76.76</td>
</tr>
<tr>
<td>Busses</td>
<td>96.00</td>
<td>94.00</td>
<td>97.50</td>
<td>94.00</td>
<td>94.07</td>
<td>92.24</td>
</tr>
<tr>
<td>Dinosaurs</td>
<td>96</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>99.02</td>
<td>99.35</td>
</tr>
<tr>
<td>Elephants</td>
<td>84.59</td>
<td>86.25</td>
<td>89.99</td>
<td>86.25</td>
<td>87.01</td>
<td>81.38</td>
</tr>
<tr>
<td>Flower</td>
<td>90.90</td>
<td>91.05</td>
<td>94.01</td>
<td>91.05</td>
<td>88.65</td>
<td>83.40</td>
</tr>
<tr>
<td>Horses</td>
<td>78.02</td>
<td>85.28</td>
<td>86.38</td>
<td>85.28</td>
<td>88.61</td>
<td>82.81</td>
</tr>
<tr>
<td>Mountain</td>
<td>87.52</td>
<td>82.36</td>
<td>82.85</td>
<td>82.36</td>
<td>79.29</td>
<td>78.60</td>
</tr>
<tr>
<td>Food</td>
<td>90.63</td>
<td>85.05</td>
<td>85.88</td>
<td>85.05</td>
<td>83.87</td>
<td>82.71</td>
</tr>
</tbody>
</table>

3.1 Average retrieval precision obtained from the proposed image representations by using DBN, SVM and RBF-NN.

According to the experimental results, the best mean average precision is obtained from the proposed image representation (Level 2) by using DBN with a value of 87.65%. The best mean average precision obtained by using SVM and RBF-NN (Level 2) is 86.27% and 84.87% respectively. The precision and recall values of the proposed image representation (Level 2) for the top 20 retrievals calculated by using DNN and SVM are compared with the results of state-of-the-art research [9, 106, 111].
Figure 3.7 Graphical representation of the mean average precision obtained from the Corel A image dataset.

Table 3.2 Class-wise comparison of average recall obtained from Corel A image dataset.
Table 3.3 Class-wise comparison of average recall obtained from Corel A image dataset.

<table>
<thead>
<tr>
<th>Class/Method</th>
<th>Level 2 DBN</th>
<th>Level 2 SVM</th>
<th>Rectangular (2x2 SVM)</th>
<th>Ashraf et al.[111]</th>
<th>Irtaza et al.[107]</th>
<th>Wang et al.[9]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>15.67</td>
<td>13.82</td>
<td>13.70</td>
<td>13.00</td>
<td>13.00</td>
<td>12.80</td>
</tr>
<tr>
<td>Beach</td>
<td>16.49</td>
<td>14.44</td>
<td>14.24</td>
<td>14.00</td>
<td>12.00</td>
<td>10.80</td>
</tr>
<tr>
<td>Building</td>
<td>18.32</td>
<td>16.97</td>
<td>16.65</td>
<td>15.00</td>
<td>12.40</td>
<td>10.60</td>
</tr>
<tr>
<td>Buses</td>
<td>19.20</td>
<td>19.50</td>
<td>19.20</td>
<td>19.00</td>
<td>17.00</td>
<td>18.80</td>
</tr>
<tr>
<td>Dinosaurs</td>
<td>19.30</td>
<td>20.00</td>
<td>20.00</td>
<td>20.00</td>
<td>18.60</td>
<td>19.60</td>
</tr>
<tr>
<td>Elephants</td>
<td>16.92</td>
<td>18.00</td>
<td>17.83</td>
<td>16.00</td>
<td>13.00</td>
<td>15.60</td>
</tr>
<tr>
<td>Flower</td>
<td>18.18</td>
<td>18.80</td>
<td>18.50</td>
<td>19.00</td>
<td>18.80</td>
<td>14.20</td>
</tr>
<tr>
<td>Horses</td>
<td>15.60</td>
<td>17.28</td>
<td>17.24</td>
<td>18.00</td>
<td>15.40</td>
<td>18.60</td>
</tr>
<tr>
<td>Mountain</td>
<td>17.50</td>
<td>16.57</td>
<td>16.28</td>
<td>15.00</td>
<td>14.60</td>
<td>8.40</td>
</tr>
<tr>
<td>Food</td>
<td>18.13</td>
<td>17.18</td>
<td>16.72</td>
<td>15.00</td>
<td>16.20</td>
<td>10.00</td>
</tr>
</tbody>
</table>

Figure 3.8 Graphical representation of the mean average precision obtained from Corel A image benchmark.
Figure 3.9 Image retrieval result obtained by using the proposed image representation (Level 2) for the class of “Busses”.

Figure 3.10 Image retrieval result obtained by using the proposed image representation (Level 2) for the class of “Buildings”.
Figure 3.11 Single label image annotation result for the semantic class “Beach”.

Figure 3.12 Single label image annotation result for the semantic class “Food”.
Figure 3.13 Two labels image annotation result for the semantic class “Dinosaurs”.

Figure 3.14 Two labels image annotation result for the semantic class “Flowers”.
Figure 3.15 Two labels image annotation result for the semantic class “Horses”.

Figure 3.16 Three labels image annotation result for the semantic class “Horses”.
3.4.2 Performance on the Fifteen Scene image benchmark

There are fifteen semantic classes in the Fifteen Scene image benchmark and each class contains 200-400 images. There are a total of 4485 images with an average resolution of 300x250 pixels. The images are divided into the semantic categories of Office, Kitchen, Living Room, Bedroom, Store, Industrial, Tall Building, Inside City, Street, Highway, Coast, Open Country, Mountain, Forest and Suburb (shown in Figure 3.17). The mean average precision values obtained from the proposed image representations are compared with rectangular histograms (2x2) and the standard BoF representation.

The results are reported for a random selection of 1352 images from different classes of test dataset. Various sizes of visual vocabularies are constructed [100, 200, 300, 400, 800, 1000, 1500] and the mean average precision on vocabularies of different sizes is calculated. The best mean average precision is obtained from the proposed image representations on a vocabulary with a size of 800, which is graphically represented in Fig.3.18. The experimental results obtained from the Fifteen Scene category image dataset and the comparisons made with them prove the robustness of the proposed image representations based on triangular histograms. The histograms of triangles (Level 2) outperform the rectangular histograms (2x2) while the histograms of triangles (Level 1) outperform the standard BoF representation. The best mean average precision obtained from the proposed image representation (Level 2) using DBN and SVM is 79.7% and 77%, respectively.
3.4.3 Run Time Analysis

We performed two types of run time analysis, in the first we presented the comparison of time in seconds required to compute the histograms of triangles from feature extraction to histogram computation (presented in table 3.4). The second analysis is presented in term of time required (in seconds) to retrieve the number of the images (presented in table 3.5, total time in seconds from feature extraction to image retrieval).

![Histogram of Triangular Regions](image)

Figure 3.18 Comparison of mean average precision using Fifteen Scene image dataset.

<table>
<thead>
<tr>
<th>Level 1 triangles</th>
<th>Level 2 triangles</th>
<th>RSHD [100]</th>
<th>Color SIFT [110]</th>
<th>CDH [100]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2875</td>
<td>0.2980</td>
<td>0.375</td>
<td>5.6</td>
<td>1.709</td>
</tr>
</tbody>
</table>

Table 3.4 Comparison of time (in seconds) for histogram computation.
3.5 Chapter Summary

In this chapter, we proposed a novel image representation that is based on spatial triangular histograms (Level 1 and Level 2) that adds spatial information to the inverted index of BoF representation. Semantic information is available at the top, right, left and bottom of the image. Constructing histograms from triangular image areas is a possible solution for reducing the semantic gap and adding the spatial attributes of images to image retrieval. The proposed image representations are evaluated by using two challenging image datasets. Three different classifiers are selected in order to evaluate the best performance of the proposed image representations. The performance of the proposed image representations is compared with existing state-of-the-art research, including rectangular histograms (2x2). The use of triangular histograms (Level 2) with DBN is found to be robust and to outperform the state-of-the-art techniques, including rectangular histograms (2x2) and the standard BoF representation.

<table>
<thead>
<tr>
<th>Number of images retrieved</th>
<th>Histograms of Triangles (Level 1)</th>
<th>Histograms of Triangles (Level 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 images</td>
<td>0.3262</td>
<td>0.3726</td>
</tr>
<tr>
<td>10 images</td>
<td>0.4579</td>
<td>0.5178</td>
</tr>
<tr>
<td>15 images</td>
<td>0.6813</td>
<td>0.7050</td>
</tr>
<tr>
<td>20 images</td>
<td>0.8423</td>
<td>0.8882</td>
</tr>
<tr>
<td>25 images</td>
<td>1.0357</td>
<td>1.0599</td>
</tr>
</tbody>
</table>

Table 3.5 Time required for retrieving images.
Chapter-4

4 Visual Words Integration of SIFT and SURF

In Content-Based Image Retrieval (CBIR), high-level visual information is represented in the form of low-level features. The semantic gap between the low-level features and the high-level image concepts is an open research problem. In this chapter, we present a novel visual words integration of Scale Invariant Feature Transform (SIFT) and Speeded-Up Robust Features (SURF). The two local features representations are selected for image retrieval because SIFT is more robust to the change in scale and rotation, while SURF is robust to changes in illumination. The visual words integration of SIFT and SURF adds the robustness of both features to image retrieval. The qualitative and quantitative comparisons conducted on Corel-1000, Corel-1500, Corel-2000, Oliva and Torralba and Ground Truth image benchmarks demonstrate the effectiveness of the proposed visual words integration. The work presented in Chapter 3 is using the BoF based image representation model. The proposed research presented in this chapter is also using the same image representation model that is using the classification framework. In addition to this the proposed research based on visual words integration of SIFT and SURF represents an image in a dimension that is twice the size of constructed vocabulary. Half of visual words represent the clusters centers of SIFT while remaining half represent the visual words of SURF.

4.1 Introduction

Color, texture and shape are examples of the global low-level features that can describe the content-based attributes of an image [3]. Color histograms are invariant to changes in scale and rotation [5]. The color features do not represent spatial distribution; moreover the closeness of the color values of two images belonging to different classes results in the output of irrelevant images [1, 3]. Texture features represent spatial variations in the group of pixels and are classified into two categories [2]. Spatial texture techniques are sensitive to noise and distortion, while spectral texture techniques work effectively on square regions by using the Fast Fourier
Visual Words Integration of SIFT & SURF

Transform (FFT) [2]. Zhang et al.[83] classify shape features into two categories: region-based and contour-based. Region-based approaches extract shape features from the entire region, and are mostly applied together with color features[83]. Contour-based approaches are applied to extract features from the edges of an image and are sensitive to noise [2].

The appearance of a similar view in images belonging to different classes, results in the closeness of the feature vector values; it also decreases the performance of image retrieval [1, 3]. The main requirement of the research in CBIR is to sort the images that are visually similar to the query image on the basis of image contents [1, 3]. Figure 4.1 represents four images of two different classes from the Corel image benchmark with a close visual similarity and semantic likeness. The human eye groups all of these images together as similar in terms of color, while at the same time recognizing a high-level semantic content. In contrast, a closer look leads to the result that the two images in the first row belong to the semantic class Mountains, while the images in the second row belong to the class Beach. While there are visual similarities like sky, clouds, people and water in both of the categories, based on the user preferences during a search, an image retrieval system must be able to retrieve images that meet the specific requirements [1, 3].

![Figure 4.1 Images of different semantic classes from the Corel image benchmark.](image)

In general, CBIR methods can be classified into two groups that employ local and global features [1, 144]. To support the visual queries, i.e. to retrieve visually similar images, mainly
Visual Words Integration of SIFT & SURF

Chapter-4

global features are used [5]. In most cases, the global features are able to capture an abstract level of semantic similarity [145]. While global features are able to identify the fact that all of the aforementioned images belong to the semantic class “natural landscapes”, usually their results are notoriously noisy [145]. By employing a global feature, a query image of a red tomato on a white background would retrieve a red pie-chart on white paper in the early positions [146]. On the other side of the spectrum, systems that support semantic queries primarily use local features, as they are able to sort the retrieved results more accurately [144, 145, 147]. If a user queries an image depicting a mountain, the retrieval system will firstly sort visually similar images that illustrate mountains (a more detailed semantic description). In the sequel, the system will include visually similar images from a higher-level semantic class. According to the recent literature, local features provide slightly better retrieval effectiveness than global features [144, 147, 148].

In recent years, local features such as SIFT [81], Histogram of Oriented Gradients (HOG) [102], SURF[82], Binary Robust Invariant Scalable Keypoints (BRISK) [149] and Maximally Stable Extremal Regions (MSER) [103] have been applied for robust content-based image matching [145, 150, 151]. There are numerous studies on local features that are associated with different applications [145].

Using local features, the representation of the image is mapped into a high-dimensional local feature space [152]. In applications such as Visual Simultaneous Localization And Mapping (VSLAM), panorama construction and object recognition, these extracted features are used directly to find one-to-one matches between depictions. In CBIR, perfect retrieval results have not been reported yet because a single feature-based image representation is not robust for all transformations [109, 119]. The visual features are combined to enhance the effectiveness and reliability of image retrieval [2, 5, 109, 119]. SIFT and SURF are reported as two robust local features and both are evaluated on different image datasets [153]. According to the experimental results[154], SIFT is more robust to rotation, change of scale, and is capable of capturing local object edges and shape by using the distribution of the intensity gradients [81]. SIFT performs accurately on the images with a simple background and represents them without noise interference [109, 119]. The performance of SIFT decreases with a complex noisy background and changes in illumination [109, 119]. SURF is reported to be robust to changes in illumination [154] and the SURF descriptor is more distinctive [155]. We show that by integrating the visual
words of SIFT and SURF, more precise, effective, and reliable image retrieval results can be obtained.

Keeping these facts in mind, this chapter, presents a novel lightweight visual words integration of SIFT and SURF. The local features are extracted from the images; for a compact representation, the feature space is quantized and two codebooks are constructed by using features of SIFT and SURF, respectively. The clusters /visual words of two local features are combined to represent image in a dimension that is twice the size of the constructed vocabulary or codebook. Here are the main contributions of this chapter:

1. Image retrieval based on visual words integration of SIFT and SURF.
2. Reduction of the semantic gap between low-level features and high-level image concepts.

4.2 Visual Words Integration of SIFT-SURF

The proposed image representation is based on the BoF representation [7]. Figure 4.2 represents the block diagram of the proposed framework. SIFT, SURF, visual words integration using the BoF representation as well as image classification are discussed in detail in the following subsections.

4.2.1 Scale invariant feature transform (SIFT)

Scale space extrema detection, keypoints localization, orientation assignment and keypoint descriptor are the four major steps for computing the SIFT descriptor [81].

1. In the first step, the Difference-of-Gaussian (DoG) is applied for the calculation of potential interest points and several Gaussian blurred images are produced by applying different scales to the input image. The DoG is calculated by using the neighborhood blur images.

\[ L(x, y, \sigma) = G(x, y, \sigma) \ast I(x, y) \]

(4.1)

Where \( L \) is the blurred image, \( I \) is the input image, \( x, y \) are the coordinates, \( \sigma \) is the scale parameter and \( \ast \) is convolution operator, the Gaussian blur operator \( G \) is:

\[
G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}
\]
A series of DoG is applied to the scale space and stable keypoints are detected by using the maxima and minima of the Laplacian of Gaussian, $\sigma^2 \nabla^2 G$, that is mathematically expressed as:

$$G(x, y, \sigma) = \frac{1}{2\pi \sigma} e^{-\frac{(x^2 + y^2)}{2\sigma^2}} \quad (4.2)$$
2. In the second step, the extrema are calculated in DoG images for the selection of candidate keypoints. Taylor series is applied to eliminate low contrast and poor localized candidate along the edges. Traces and determinant of Hessian metrics are calculated and thresholding is used to remove the outliers.

\[
D(X) = D + \frac{\partial D^T}{\partial x} x + \frac{1}{2} x^T \frac{\partial^2 D}{\partial x^2} x
\]  

where \(D\) and its derivate are evaluated at point \(x = (x, y, \sigma)^T\).

3. In the third step, the principal orientation is assigned to the keypoints and achieves invariance to image rotation. The magnitude and the direction of the keypoints are expressed by the following equation:

\[
m(x, y) = \sqrt{(L(x + 1, y) - L(x - 1, y))^2 + (L(x, y + 1) - L(x, y - 1))^2} \quad (4.4)
\]

\[
\theta(x, y) = \tan^{-1}((L(x, y + 1) - L(x, y - 1))(L(x + 1, y) - L(x - 1, y)) \quad (4.5)
\]

4. The fourth step computes the SIFT descriptor across each keypoint. The descriptor gradient orientations and coordinates are rotated relative to the keypoint orientation and provide the orientation invariance. For each keypoint, a set of orientation histograms are created on 4x4 pixel neighborhood, with 8 orientation bins in each. This result in a feature vectors containing 128 dimensions, SIFT descriptors are invariant to contrast, scale and rotation [81].

### 4.2.2 Speeded-up robust features (SURF)

There are two main steps to compute the SURF keypoints and descriptors [82].

1. The box filter is applied to the integral images for an efficient computation of the Laplacian of Gaussian. Determinants of the Hessian matrix are calculated for the detection of the keypoints.
\[ H(X, \sigma) = \begin{bmatrix} L_{xx}(X, \sigma) & L_{xy}(X, \sigma) \\ L_{xy}(X, \sigma) & L_{yy}(X, \sigma) \end{bmatrix} \] (4.6)

\[ L_{xx}(X, \sigma) = I(x, y) \otimes \frac{\partial^2}{\partial x^2} g(\sigma) \] (4.7)

where \( X = I(x, y) \) is the image and \( L_{xx}(x, \sigma) \) is the convolution of the Gaussian second order derivative with a scale \( \sigma \) at the point \( X \), while \( L_{xy}(x, \sigma) \) and \( L_{yy}(x, \sigma) \) are representing the convolution of the second order derivative \( \frac{\partial^2}{\partial x^2} g(\sigma) \) of the input image \( X \), while \( xy \) and \( yy \) are the diagonal and vertical directions respectively. The scale space in SURF is constructed by varying the filter size and this result in calculating the scale space that is invariant to the location and scale.

2. In the second step, every keypoint is assigned to a reproducible orientation by applying the Haar wavelet in the direction of \( x \) and \( y \). A square window is applied around the keypoints and is oriented along the orientations detected before. The Haar wavelets with a size of 2 sigma are calculated by applying the window that is divided into 4x4 regular sub-regions and each sub-region contributes values. This result in a feature vector containing 64 dimensions, SURF descriptors are invariant to rotation, change of scale and contrast [82].

### 4.2.3 Visual words integration using BoF model

The proposed image representation is based on the visual words integration of SIFT and SURF by using the BoF representation [7].

1. The BoF representation an image \( R \) is represented as:

\[ R = (a_{m,n}) \] (4.8)

Where \( m \) and \( n \) are the co-ordinates along \( x \) and \( y \) axis respectively.

2. Local features (SIFT & SURF) are extracted from an image \( R \) is represented as:
\[ R = \{d_1, d_2, d_3, \ldots, d_M\} \quad (4.9) \]

Where \(d_1\) to \(d_M\) are the descriptors of image \(R\) that are extracted by applying the detectors of SIFT & SURF, respectively.

3. A clustering algorithm such as \(k\)-means [7] is applied to reduce/quantize the feature space into \(t\) number of words that are represented as

\[ voc = \{v_1, v_2, v_3, \ldots, v_t\} \quad (4.10) \]

Where \(v_1\) to \(v_t\) represents the visual words (or clusters) while \(voc\) represents the constructed visual vocabulary (codebook with \(t\) clusters). Separate visual vocabulary is constructed by using the features of SIFT and SUR, respectively.

4. To represent an image in the form of visual words by using the constructed visual vocabulary, local features are extracted and quantized in the feature space. The following equation is used to assign the visual words to the reduced/quantized feature space.

\[ v(d_k) = \arg\min_{v \in voc} Dist(v, d_k) \quad (4.11) \]

Where \(v(d_k)\) is the visual word that is assigned to the \(k^{th}\) descriptor \(d_k\) and the Euclidean distance between the feature space and visual word is represented by the \(Dist(v, d_k)\).

Visual words of SIFT are mapped on images by using the vocabulary of SIFT and visual words of SURF are mapped on the images by the use of vocabulary of SURF.

5. The visual words of SIFT and SURF are vertically concatenated to represent an image in the form of visual words of two different local features. The visual word integration of SIFT and SURF represents an image in a dimension that is twice of the size of the constructed visual vocabulary.
4.3 Image classification framework

Support Vector Machine (SVM) works on the basis of supervised classification method [2]. Two classes are separated in a linear SVM with the help of a hyperplane. The classification case of two classes can be mathematically expressed as:

\[ \{(x_i + y_i)\}^N_{i=1}y_i = \{+1, -1\} \]  \hspace{1cm} (4.12)

Where the data of two classes are expressed with \(x_i\) and \(y_i\) respectively while the label are +1 and +1. The following coefficients are determined to find the hyperplane.

\[ w^T x + b = 0 \]  \hspace{1cm} (4.13)

In this case \(w\) represents the weight vector while \(b\) is bias. The margin is calculated by using \(2/\|w\|\). Following equation is applied to separate the classes:

\[ w^T x + b = 1 \]  \hspace{1cm} (4.14)

\[ w^T x + b = -1 \]  \hspace{1cm} (4.15)

The relevant detail about the parameter for SVM optimization is mentioned in chapter 3, section 3.3.1.

4.4 Experiments and Results

The details about the experimental parameters that are used for the evaluation of proposed research are mentioned in this section. The proposed image representation is evaluated on Corel-1000 [143], Corel-1500 [157], Corel-2000 [157], Oliva and Torralba [158] and GT image benchmarks. SIFT and SURF are used for features extraction, and therefore all of the images are processed in gray scale. We applied \(k\)-means unsupervised clustering to build/construct the codebook therefore, we repealed all experiments 10 times and the average/mean values are calculated and reported in the tables and graphs. For every experiment, training and test datasets are selected randomly. The performance of image retrieved based on BoF model depends on the visual vocabulary size or number of cluster of the codebook [122, 125]. An increase in the
vocabulary size/number of clusters in the codebook results an increase in the mean performance while on larger vocabulary size there is a decrease in mean performance due to over-fitting [125]. We constructed different sizes of visual vocabulary/number of clusters from a set of training images to sort out/ determine the best performance of the proposed research. Different number of features percentages per image from the training dataset is used to construct the visual vocabulary to validate the mean retrieval performance. The images are processed in the standard size as available in the respective dataset.

### 4.4.1 Weighted average of SIFT and SURF

The proposed image representation is based on the visual words integration of SIFT and SURF. Differently weighted averages of SIFT and SURF are also calculated to report the second best retrieval performance. The weighted average (WA) of SIFT and SURF is calculated by using the following equation:

\[
WA = \frac{w \cdot FV_{SIFT} + (1 - w) \cdot FV_{SURF}}{2}
\]  

(4.16)

Where \(FV_{SIFT}\) and \(FV_{SURF}\) are the feature vectors consisting of visual words of SIFT and SURF respectively and \(0 < w < 1\).

The weighted average of SIFT and SURF represents an image in a dimension that is equal to the size of the constructed visual vocabulary.
Figure 4.3 Samples of images from each category of the Corel-1000 image benchmark [143].

### 4.4.2 Performance using Corel-1000 image benchmark

The Corel-1000 image benchmark is a sub-set of the Corel image dataset and is extensively used to evaluate CBIR research [9, 110, 119]. The Corel-1000 image benchmark contains 1000 images divided into 10 semantic classes. Figure 4.3 represents the images from all of the categories from the Corel-1000 image benchmark. The Corel-1000 image benchmark is selected for the evaluation of the proposed image representation and image retrieval precision is compared with existing state-of-the-art CBIR approaches. Testing is performed by a random selection of 500 images from the test dataset. The mean average precision of the proposed image representation is evaluated by using different sizes of vocabulary [50, 100, 200, 400, 600, 800, 1000, 1200]. Different weighted averages of SIFT and SURF are also calculated to find out the second best performance on the Corel-1000 image benchmark. The weighted average values used in the experimental work for SIFT-SURF are 1.0-0.0, 0.9-0.1, 0.8-0.2, 0.7-0.3, 0.6-0.4, 0.5-0.5, 0.4-0.6, 0.3-0.7, 0.2-0.8, 0.1-0.9 and 0.0-1.0, where the first value represents the weight of SIFT and the second value represents the weight of SURF. The best mean average precision is obtained when using the weighted average of 0.7-0.3 (SIFT-SURF). The mean average precision, sigma, and confidence interval (CI) for the top
20 retrievals obtained by using visual words integration and weighted average of 0.7-0.3 (SIFT-SURF) is represented in Table 4.1 and Table 4.2, respectively.

<table>
<thead>
<tr>
<th>Vocabulary size and features % used</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>400</th>
<th>600</th>
<th>800</th>
<th>1000</th>
<th>1200</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>68.87</td>
<td>71.01</td>
<td>74.05</td>
<td>74.46</td>
<td>74.94</td>
<td>74.54</td>
<td>74.77</td>
<td>74.67</td>
</tr>
<tr>
<td>25%</td>
<td>69.62</td>
<td>72.52</td>
<td>74.36</td>
<td>74.75</td>
<td>75.48</td>
<td>75.02</td>
<td>75.2</td>
<td>75.53</td>
</tr>
<tr>
<td>50%</td>
<td>68.4</td>
<td>72.13</td>
<td>73.62</td>
<td>75.27</td>
<td>75.55</td>
<td>75.4</td>
<td>75.72</td>
<td>75.4</td>
</tr>
<tr>
<td>75%</td>
<td>69.94</td>
<td>71.87</td>
<td>73.24</td>
<td>75.78</td>
<td>74.75</td>
<td>75.53</td>
<td>74.75</td>
<td>74.84</td>
</tr>
<tr>
<td>100%</td>
<td>69.87</td>
<td>72.69</td>
<td>74.23</td>
<td>74.04</td>
<td>75.15</td>
<td>75.05</td>
<td>74.62</td>
<td>74.89</td>
</tr>
<tr>
<td>Mean</td>
<td>69.34</td>
<td>72.04</td>
<td>73.9</td>
<td>74.86</td>
<td>75.17</td>
<td>75.10</td>
<td>75.01</td>
<td>75.06</td>
</tr>
<tr>
<td>CI</td>
<td>±0.84</td>
<td>±0.82</td>
<td>±0.57</td>
<td>±0.85</td>
<td>±0.42</td>
<td>±0.47</td>
<td>±0.56</td>
<td>±0.46</td>
</tr>
<tr>
<td>Sigma</td>
<td>0.68</td>
<td>0.66</td>
<td>0.46</td>
<td>0.68</td>
<td>0.34</td>
<td>0.73</td>
<td>0.45</td>
<td>0.38</td>
</tr>
</tbody>
</table>

Table 4.1 Mean average precision for top 20 retrievals (visual words integration).

<table>
<thead>
<tr>
<th>Vocabulary size and features % used</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>400</th>
<th>600</th>
<th>800</th>
<th>1000</th>
<th>1200</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>61.95</td>
<td>66.39</td>
<td>67.84</td>
<td>68.89</td>
<td>66.99</td>
<td>70.25</td>
<td>70.38</td>
<td>69.73</td>
</tr>
<tr>
<td>25</td>
<td>62.45</td>
<td>66.86</td>
<td>67.56</td>
<td>67.73</td>
<td>70.2</td>
<td>71.2</td>
<td>70.6</td>
<td>70.21</td>
</tr>
<tr>
<td>50</td>
<td>61.69</td>
<td>67.45</td>
<td>66.55</td>
<td>69.17</td>
<td>69.84</td>
<td>70.39</td>
<td>69.9</td>
<td>69.35</td>
</tr>
<tr>
<td>75</td>
<td>62.03</td>
<td>66.49</td>
<td>67.24</td>
<td>68.84</td>
<td>69.65</td>
<td>70.5</td>
<td>70.15</td>
<td>70.01</td>
</tr>
<tr>
<td>100</td>
<td>62.09</td>
<td>66.37</td>
<td>67.96</td>
<td>68.85</td>
<td>69.95</td>
<td>70.57</td>
<td>70.1</td>
<td>70.2</td>
</tr>
<tr>
<td>Mean</td>
<td>62.04</td>
<td>66.71</td>
<td>67.43</td>
<td>68.9</td>
<td>69.93</td>
<td>70.58</td>
<td>70.23</td>
<td>69.9</td>
</tr>
<tr>
<td>CI</td>
<td>±0.34</td>
<td>±0.57</td>
<td>±0.7</td>
<td>±0.95</td>
<td>±0.25</td>
<td>±0.45</td>
<td>±0.34</td>
<td>±0.45</td>
</tr>
<tr>
<td>Sigma</td>
<td>0.274</td>
<td>0.457</td>
<td>0.565</td>
<td>0.763</td>
<td>0.202</td>
<td>0.366</td>
<td>0.269</td>
<td>0.363</td>
</tr>
</tbody>
</table>

Table 4.2 Mean average precision for top 20 retrievals (weighted average 0.7-0.3).

According to the experimental results obtained by applying the visual words integration of SIFT and SURF, the best mean average precision of 75.17% is obtained on a vocabulary with a size 600 (by calculating the mean of all columns on the vocabulary of a size of 600 in Table 4.1). Table 4.2 represents the mean average precision obtained from the weighted average of 0.7-0.3 (SIFT-SURF). The best mean average precision of 70.58% is obtained on a vocabulary with a size of 800 (by calculating the mean of all columns of the vocabulary of a size of 600 in Table...
4.2). Figure 4.4 represents the comparison of mean average precision for top 20 retrievals using visual words integration and different weighted averages.

The experimental results and comparisons conducted using the Corel-1000 image benchmark prove the robustness of the proposed image representation based on the visual words integration of SIFT and SURF. The mean average precision value obtained from the proposed framework is higher than that of the existing state-of-the-art research. Figure 4.5 represents precision-recall curve obtained using Corel-1000 image benchmark. The image retrieval results obtained from the proposed framework are represented in Figure 4.6 to Figure 4.9. The single image displayed in the first row is the query image, and the numerical value displayed at the top of each image is the classifier decision value (score) of the respective image.

![Figure 4.4 Comparison of mean average precision for top 20 retrievals using the Corel-1000.](image-url)
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>60.08</td>
<td>52.68</td>
<td>74.6</td>
<td>64</td>
<td>57</td>
<td>55</td>
</tr>
<tr>
<td>Beach</td>
<td>60.39</td>
<td>56.32</td>
<td>37.8</td>
<td>54</td>
<td>58</td>
<td>47</td>
</tr>
<tr>
<td>Building</td>
<td>69.66</td>
<td>69.03</td>
<td>53.9</td>
<td>53</td>
<td>43</td>
<td>56</td>
</tr>
<tr>
<td>Buses</td>
<td>93.65</td>
<td>86.35</td>
<td>96.7</td>
<td>94</td>
<td>93</td>
<td>91</td>
</tr>
<tr>
<td>Dinosaurs</td>
<td>99.88</td>
<td>99.68</td>
<td>99</td>
<td>98</td>
<td>98</td>
<td>94</td>
</tr>
<tr>
<td>Elephants</td>
<td>70.76</td>
<td>67.55</td>
<td>65.9</td>
<td>78</td>
<td>58</td>
<td>49</td>
</tr>
<tr>
<td>Flowers</td>
<td>88.37</td>
<td>85.99</td>
<td>91.2</td>
<td>71</td>
<td>83</td>
<td>85</td>
</tr>
<tr>
<td>Horses</td>
<td>82.77</td>
<td>76.37</td>
<td>86.9</td>
<td>93</td>
<td>68</td>
<td>52</td>
</tr>
<tr>
<td>Mountains</td>
<td>61.08</td>
<td>58.85</td>
<td>58.5</td>
<td>42</td>
<td>46</td>
<td>37</td>
</tr>
<tr>
<td>Food</td>
<td>65.09</td>
<td>53.00</td>
<td>62.2</td>
<td>50</td>
<td>53</td>
<td>55</td>
</tr>
<tr>
<td>Mean</td>
<td>75.17</td>
<td>70.58</td>
<td>72.67</td>
<td>69.7</td>
<td>65.7</td>
<td>62.1</td>
</tr>
</tbody>
</table>

Table 4.3 Class-wise comparison of precision for top 20 retrievals.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>12.02</td>
<td>10.54</td>
<td>14.92</td>
<td>12.80</td>
<td>11.40</td>
<td>11.00</td>
</tr>
<tr>
<td>Beach</td>
<td>12.08</td>
<td>11.26</td>
<td>7.56</td>
<td>10.80</td>
<td>11.60</td>
<td>9.40</td>
</tr>
<tr>
<td>Building</td>
<td>13.93</td>
<td>13.81</td>
<td>10.78</td>
<td>10.60</td>
<td>8.60</td>
<td>11.20</td>
</tr>
<tr>
<td>Buses</td>
<td>18.73</td>
<td>17.27</td>
<td>19.34</td>
<td>18.80</td>
<td>18.60</td>
<td>18.20</td>
</tr>
<tr>
<td>Dinosaurs</td>
<td>19.98</td>
<td>19.94</td>
<td>19.80</td>
<td>19.60</td>
<td>19.60</td>
<td>18.80</td>
</tr>
<tr>
<td>Elephants</td>
<td>14.15</td>
<td>13.51</td>
<td>13.18</td>
<td>15.60</td>
<td>11.60</td>
<td>9.80</td>
</tr>
<tr>
<td>Flowers</td>
<td>17.67</td>
<td>17.20</td>
<td>18.24</td>
<td>14.20</td>
<td>16.60</td>
<td>17.00</td>
</tr>
<tr>
<td>Horses</td>
<td>16.55</td>
<td>15.27</td>
<td>17.38</td>
<td>18.60</td>
<td>13.60</td>
<td>10.40</td>
</tr>
<tr>
<td>Mountains</td>
<td>12.22</td>
<td>11.77</td>
<td>11.70</td>
<td>8.40</td>
<td>9.20</td>
<td>7.40</td>
</tr>
<tr>
<td>Food</td>
<td>13.02</td>
<td>10.60</td>
<td>12.44</td>
<td>10.00</td>
<td>10.60</td>
<td>11.00</td>
</tr>
<tr>
<td>Mean</td>
<td>15.03</td>
<td>14.12</td>
<td>14.53</td>
<td>13.94</td>
<td>13.14</td>
<td>12.42</td>
</tr>
</tbody>
</table>

Table 4.4 Class-wise comparison of recall for top 20 retrievals.
Figure 4.5 Precision-recall curve obtained using the Corel-1000 image benchmark.

Figure 4.6 Image retrieval results for the class Horses.
Figure 4.7 Image retrieval results for the class Dinosaurs.

Figure 4.8 Image retrieval results for the class Elephants.
4.4.3 Performance on the Corel-1500 image benchmark

The Corel-1500 image benchmark contains 1500 images (divided into 15 semantic classes) and is a sub-set of the Corel image dataset [157]. Figure 4.10 represents the images from all of the categories from the Corel-1500 image benchmark. Testing is performed by a random selection of
750 images from the test dataset. Figure 4.11 represents the comparison of mean average precision using visual words integration and different weighted averages.

![Figure 4.11](image_url)

Figure 4.11 Comparison of precision and recall using the Corel-1500 image benchmark.

According to the experimental results, the best mean average precision obtained from the visual words integration of SIFT and SURF on a vocabulary with a size of 600 is 74.95%. The best mean average precision obtained using the weighted average of 0.7-0.3 (SIFT-SURF) on a vocabulary with a size of 800 is 68.05%. The visual words integration of SIFT and SURF significantly enhances the performance of image retrieval. The comparison of precision and recall obtained from the proposed framework and state-of-the-art research [159] is presented in Table 4.5.

<table>
<thead>
<tr>
<th>Performance/Method</th>
<th>Visual words integration</th>
<th>Weighted average 0.7-0.3</th>
<th>SQ+Spatiogram[159]</th>
<th>GMM+mSpatiogram[159]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>74.95±1.60</td>
<td>68.05±1.92</td>
<td>63.95</td>
<td>74.10</td>
</tr>
<tr>
<td>Recall</td>
<td>14.99±0.32</td>
<td>13.15±0.38</td>
<td>12.79</td>
<td>13.80</td>
</tr>
</tbody>
</table>

Table 4.5 Comparison of precision and recall using the Corel-1500 image benchmark.
4.4.4 Performance on the Corel-2000 image benchmark

The Corel-2000 image benchmark contains 2000 images (divided into 20 semantic classes) and is a sub-set of Corel image dataset. Figure 4.12 represents the images from all of the categories from the Corel-2000 image benchmark. Testing is performed by a random selection of 600 images from the test dataset. Figure 4.13 represents the comparison of mean average precision using visual words integration and different weighted averages.

According to the experimental results, the best mean average precision obtained from the visual words integration of SIFT and SURF on a vocabulary with a size of 800 is 65.41%. The best mean average precision of 58.31% is obtained when using the weighted average of 0.3-0.7 (SIFT-SURF). The visual words integration of SIFT and SURF significantly enhances the performance of image retrieval. The comparison of the mean average precision obtained from the proposed frame work and state-of-the-art research [160, 161] is presented in Table 4.6.
Figure 4.13 Comparison of precision and recall using the Corel-2000 image benchmark.

<table>
<thead>
<tr>
<th>Performance/Method</th>
<th>Visual words integration</th>
<th>MissSVM [160]</th>
<th>MI-SVM [161]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>65.41±0.99</td>
<td>65.2</td>
<td>54.6</td>
</tr>
</tbody>
</table>

Table 4.6 Comparison of precision and recall using the Corel-2000 image benchmark.

4.4.5 Performance on the Oliva and Torralba (OT-Scene) image benchmark

Figure 4.14 Samples of images from each category of the OT-Scene image benchmark [158].
The Oliva and Torralba (OT-Scene) image benchmark [158] was created by MIT and there are 2688 images that are divided into 08 classes. Figure 4.14 represents the images from all of the categories from the OT-Scene image benchmark. Testing is performed by a random selection of 600 images from the test dataset. Figure 4.15 represents the comparison of mean average precision using visual words integration and different weighted averages. The comparison of the mean average precision obtained from the proposed framework and state-of-the-art CBIR research [162, 163] is presented in Table 4.7.

According to the experimental results, the best mean average precision obtained using visual words integration and weighted average of 0.3-0.7 (SIFT-SURF) is 69.75% and 65.25%, respectively. The visual words integration of SIFT and SURF significantly enhances the performance of image retrieval.
Performance/Method | Visual words integration | Weighted average 0.3-0.7 | Feature extraction with morphological operators[162] | Min Max fusion [163]  
--- | --- | --- | --- | ---  
Mean | 69.75±0.40 | 65.25±0.52 | 60.7 | 51.04  

Table 4.7 Comparison of mean average precision using OT-Scene image benchmark.

### 4.4.6 Performance on the GT image benchmark

![Figure 4.16](image1.png)

Figure 4.16 Samples of images from 05 classes of GT image benchmark used for the evaluation of the proposed visual word integration.

GT image benchmark was created by University of Washington and it contains 22 classes. In order to perform a clear comparison with existing research, we selected 05 classes that are Arbor Greens, Cherries, Football, Green Lake and Swiss Mountains (shown in Figure 4.16). Different sizes of the visual vocabulary are constructed from the training dataset (10, 20, 50, 75, 100) to
sort out the best performance of the proposed framework. The best mean average precision is obtained on a vocabulary with a size of 75 with a value of 83.53%. The comparison of the mean average precision obtained from the proposed framework and existing state-of-the-art research [118, 164, 165] is presented in Table 4.8.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>83.53±1.50</td>
<td>81.33</td>
<td>62.80</td>
<td>59.09</td>
</tr>
</tbody>
</table>

Table 4.8 Mean average precision comparison by selecting 05 classes of GT image benchmark.

The numerical values of experimental results presented in Table 4.8 reflect the better performance of proposed visual word integration. The mean precision of the proposed visual words integration is higher than the research that is selected for comparison.

### 4.5 Chapter Summary

The semantic gap between low-level visual features and high-level image concepts is a challenging research problem of CBIR. SIFT and SURF are reported as two robust local features and the integration of visual words of SIFT and SURF adds the robustness of both features to image retrieval. As shown by the experimental results, the proposed image representation demonstrates an impressive performance and can be safely recommended as a preferable method for image retrieval tasks. It is safe to conclude that depending on the image collection, the visual words integration of SIFT and SURF can yield good retrieval performance with the additional benefits of fast indexing and scalability. In future, we plan to evaluate our framework for large scale image retrieval (ImageNet or Flicker) by replacing SVM with state-of-the-art classification technique such as deep learning.
Chapter 5

5 Conclusion and Future Work

This chapter summarizes about the main contributions of this thesis. The limitations of the proposed research and possible future extensions are also mentioned in the following subsections.

5.1 Conclusion

The commonly used model for image retrieval is based on classification framework that learns from the training data (that are referred as training images). Different numbers of training images are used to train the classifier learning model [166]. The training data is used to manage the image variations that occur because of change of viewpoint, rotation, changes in illumination, change in scale and spatial information. There are many cases in which the large amount of training data is not enough to manage these variations. In recent few years, local image descriptors have shown a remarkable performance to the variations like change of scale, rotation and viewpoint. The local features work efficiently in image retrieval applications, as they are calculated on smaller patches. The visual word integration (late fusion) is used to enhance the performance of image retrieval. In image classification and retrieval problems, the spatial information among different regions and robust visual words integration is required to manage the above mentioned variations.

Consequently, in this thesis, two new image representations are proposed that are based on image spatial attributes and visual words integration of two local features. The proposed image representations are evaluated by using six challenging image benchmarks and results are compared with state-of-the-art recent CBIR research.

The image representation proposed in Chapter 3 is capturing the spatial attributes of image. Standard images are constructed by following the rule of thirds, which divides an image into nine equal grids; objects or regions of interest are placed at intersecting lines of the grid (either on the left or on the right). The rule of thirds represents compositional and content-based attributes of the image. We incorporated these spatial attributes into image retrieval system to enhance the
performance. In a standard image, the sky, the sun or clouds are located at the top, objects of interest are at the right or left, and the ground, grass or water are likely to be located at the bottom of the image. This sequence represents a triangular relationship between objects and regions of interest in a scene. The construction of histograms from triangular regions of the image adds the discriminating information to image retrieval in the form of objects and regions of interest that are located at the top, left, right and bottom of the image. A novel image representation is proposed by dividing an image into two (Level 1) and four (Level 2) triangular regions. Both Level 1 and Level 2 histograms of triangular regions are evaluated separately. Mapping of visual words over the divided triangular regions increases the number of bins of histogram that are used for image representation. Level 1 and Level 2 triangles represent an image in a dimension that is twice and four times the size of the construed vocabulary. The increase in dimensions for image representations decreases the computational complexity that is associated with vocabulary construction/training and it provides more data instances at the time of classifier training. The enable a higher retrieval performance by constructing a smaller vocabulary size that significantly reduces the training cost of classifier and clustering.

Histogram of triangles Level 2 outperforms state-of-the-art recent research of CBIR including histograms of rectangular regions (2x2).

The image representation proposed in Chapter 4 is based on the visual words integration of SIFT and SURF. A single feature based image representation is not robust to all of the transformations. The main purpose of selection of visual words integration is to enhance the performance of image retrieval. In computer vision, two approaches are used for features integration. The first approach is known as early fusion and in this case, the feature space is combined and single visual vocabulary is constructed that consists of both features. In the second case, separate visual vocabulary is constructed by using both features separately and visual words constructed through the vocabulary of A and B are concatenated to represent an image in the form of visual words of both features. This is also known as late fusion or visual words integration and it is reported robust in the literature as it represents an image in a dimension that is double than the size of vocabulary. The selection of appropriate local features for visual words integration is a challenging task as improper selection of local features for visual words integration decreases the performance of image retrieval. In general, CBIR methods can be classified into two groups that employ local and global features. To support the visual
queries, i.e. to retrieve visually similar images, mainly global features are used in most of the cases; the global features are able to capture an abstract level of semantic similarity. By employing a global feature, a query image of a red tomato on a white background would retrieve a red pie-chart on white paper in the early positions. On the other side of the spectrum, systems that support semantic queries primarily use local features, as they are able to sort the retrieved results more accurately. According to the recent literature, local features provide slightly better retrieval effectiveness than global features. Keeping these facts in view, we selected SIFT and SURF for visual words integration because SIFT is more robust to the change in scale and rotation, while SURF is robust to changes in illumination. The main benefits obtained by applying a visual words integration of SIFT and SURF are the improvement in performance of image retrieval and reduction in semantic gap between low-level visual features and high-level image concepts. According to the experimental results of chapter 5, the proposed image representation demonstrates an impressive performance and can be safely recommended as a preferable method for image retrieval tasks. It is safe to conclude that depending on the image collection, the visual words integration of SIFT and SURF can yield good retrieval performance with the additional benefits of fast indexing and scalability.

### 5.2 Limitations

The proposed image representations contribute to the field of computer vision and image processing. The limitations of the proposed research are explained below:

1. The visual vocabulary is constructed by applying $k$-means clustering (hard clustering) and in this case same cluster can contain non-similar features. This affects the discriminating nature of the constructed visual vocabulary as well as the image representation that is based on histograms of visual words.

2. The proposed image retrieval is using image classification framework based on BoF representation. Training of classifier requires sufficient number of images for training. The output of image retrieval is depends on the training data. The retrieval performance can be degraded in the cases where the training data is not sufficient.
5.3 Future Directions

1. The proposed image representations are using $k$-means (unsupervised clustering) for vocabulary construction. Features are considered to be identical when they are assigned the same words from the vocabulary. $K$-means is considered as hard clustering [132] and features are assigned to the cluster centers by calculating the Euclidean distance between the features and quantized descriptor. The similarity with the cluster center is directly proportional to the distance and this affects the discriminating nature of visual vocabulary. Soft clustering is reported more robust [131] than hard clustering as it assigns features to different clusters in a weighted manner. The possible future extension of the proposed representations can be the use of soft clustering for the construction of visual vocabulary.

2. The proposed image representations are not rotation invariant. The geometric relationship between the visual words within different triangular regions will be explored in the near future to obtain an image representation that will be robust to the changes in rotation.

3. The vocabulary/codebook used in the proposed framework does not contain the spatial information among visual words. In future, the proposed work can be extended by constructing the visual vocabulary with triangular features.

4. The proposed research is evaluated on smaller image benchmarks. The performance evaluation of proposed research on larger image benchmarks such as Flicker, CIFAR-10, CIFAR-100 or ImageNet by applying deep learning is a possible future extension.
References


References


References


