Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

ABSTRACT

INVESTIGATION OF NEW FEATURE DESCRIPTORS FOR IMAGE SEARCH AND CLASSIFICATION

by
Atreyee Sinha

Content-based image search, classification and retrieval is an active and important research area due to its broad applications as well as the complexity of the problem. Understanding the semantics and contents of images for recognition remains one of the most difficult and prevailing problems in the machine intelligence and computer vision community. With large variations in size, pose, illumination and occlusions, image classification is a very challenging task. A good classification framework should address the key issues of discriminatory feature extraction as well as efficient and accurate classification. Towards that end, this dissertation focuses on exploring new image descriptors by incorporating cues from the human visual system, and integrating local, texture, shape as well as color information to construct robust and effective feature representations for advancing content-based image search and classification.

Based on the Gabor wavelet transformation, whose kernels are similar to the 2D receptive field profiles of the mammalian cortical simple cells, a series of new image descriptors is developed. Specifically, first, a new color Gabor-HOG (GHOG) descriptor is introduced by concatenating the Histograms of Oriented Gradients (HOG) of the component images produced by applying Gabor filters in multiple scales and orientations to encode shape information. Second, the GHOG descriptor is analyzed in six different color spaces and grayscale to propose different color GHOG descriptors, which are further combined to present a new Fused Color GHOG (FC-GHOG) descriptor. Third, a novel Gabor-PHOG (GPHOG) descriptor is proposed which improves upon the Pyramid Histograms of Oriented Gradients (PHOG) descriptor, and subsequently a new FC-GPHOG descriptor is constructed by combining the multiple color GPHOG descriptors and employing the

Principal Component Analysis (PCA). Next, the Gabor-LBP (GLBP) is derived by accumulating the Local Binary Patterns (LBP) histograms of the local Gabor filtered images to encode texture and local information of an image. Furthermore, a novel Gabor-LBP-PHOG (GLP) image descriptor is proposed which integrates the GLBP and the GPHOG descriptors as a feature set and an innovative Fused Color Gabor-LBP-PHOG (FC-GLP) is constructed by fusing the GLP from multiple color spaces. Subsequently, The GLBP and the GHOG descriptors are then combined to produce the Gabor-LBP-HOG (GLH) feature vector which performs well on different object and scene image categories. The six color GLH vectors are further concatenated to form the Fused Color GLH (FC-GLH) descriptor. Finally, the Wigner based Local Binary Patterns (WLBP) descriptor is proposed that combines multi-neighborhood LBP, Pseudo-Wigner distribution of images and the popular bag of words model to effectively classify scene images.

To assess the feasibility of the proposed new image descriptors, two classification methods are used: one method applies the PCA and the Enhanced Fisher Model (EFM) for feature extraction and the nearest neighbor rule for classification, while the other method employs the Support Vector Machine (SVM). The classification performance of the proposed descriptors is tested on several publicly available popular image datasets. The experimental results show that the proposed new image descriptors achieve image search and classification results better than or at par with other popular image descriptors, such as the Scale Invariant Feature Transform (SIFT), the Pyramid Histograms of visual Words (PHOW), the Pyramid Histograms of Oriented Gradients (PHOG), the Spatial Envelope (SE), the Color SIFT four Concentric Circles (C4CC), the Object Bank (OB), the Context Aware Topic Model (CA-TM), the Hierarchical Matching Pursuit (HMP), the Kernel Spatial Pyramid Matching (KSPM), the SIFT Sparse-coded Spatial Pyramid Matching (Sc-SPM), the Kernel Codebook (KC) and the LBP.

INVESTIGATION OF NEW FEATURE DESCRIPTORS FOR IMAGE SEARCH AND CLASSIFICATION

by
Atreyee Sinha

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Computer Science

Department of Computer Science

May 2014

APPROVAL PAGE

INVESTIGATION OF NEW FEATURE DESCRIPTORS FOR IMAGE SEARCH AND CLASSIFICATION

Atreyee Sinha

Dr. Chengjun Liu, Dissertation Advisor — Date
Associate Professor of Computer Science, NJIT

Dr. James Geller, Committee Member — Date
Professor and Chair of Computer Science, NJIT

Dr. Durgamadhab Misra, Committee Member — Date
Professor of Electrical and Computer Engineering, NJIT

Dr. David Nassimi, Committee Member — Date
Associate Professor of Computer Science, NJIT

Dr. Vincent Oria, Committee Member — Date
Associate Professor of Computer Science, NJIT

BIOGRAPHICAL SKETCH

Author: Atreyee Sinha

Degree: Doctor of Philosophy

Date: May 2014

Undergraduate and Graduate Education:

- Doctor of Philosophy in Computer Science,
  New Jersey Institute of Technology, Newark, New Jersey, 2014
- Bachelor of Technology in Computer Science and Engineering,
  St. Thomas College of Engineering and Technology,
  West Bengal University of Technology, Kolkata, India, 2010

Major: Computer Science

Publications:

A. Sinha, S. Banerji, and C. Liu. New Color GPHOG Descriptors for Object and Scene Image Classification. *Machine Vision and Applications*, 25(2):361-375, 2014.

A. Sinha, S. Banerji, and C. Liu. Scene Image Classification using a Wigner-Based Local Binary Patterns Descriptor. In *International Joint Conference on Neural Networks*, Beijing, China, July 6-11, 2014.

S. Banerji, A. Sinha, and C. Liu. HaarHOG: Improving the HOG Descriptor for Image Classification. In *IEEE International Conference on Systems, Man, and Cybernetics*, Manchester, UK, October 13-16, 2013.

S. Banerji, A. Sinha, and C. Liu. A New Bag of Words LBP (BoWL) Descriptor for Scene Image Classification. In *15th International Conference on Computer Analysis of Images and Patterns*, York, UK, August 27-29, 2013.

S. Banerji, A. Sinha, and C. Liu. New Image Descriptors Based on Color, Texture, Shape, and Wavelets for Object and Scene Image Classification. *Neurocomputing*, 117:173-185, 2013.

A. Sinha, S. Banerji, and C. Liu. Novel Color Gabor-LBP-PHOG (GLP) Descriptors for Object and Scene Image Classification. In *The Eighth Indian Conference on Vision, Graphics and Image Processing*, Mumbai, India, December 16-19, 2012.

A. Sinha, S. Banerji, and C. Liu. Gabor-Based Novel Local, Shape and Color Features for Image Classification. In *The 19th International Conference on Neural Information Processing*, Doha, Qatar, November 12-15, 2012.

A. Sinha, S. Banerji, and C. Liu. Novel Gabor-PHOG Features for Object and Scene Image Classification. In *Structural, Syntactic, and Statistical Pattern Recognition*, Lecture Notes in Computer Science, volume 7626, pages 584-592, 2012.

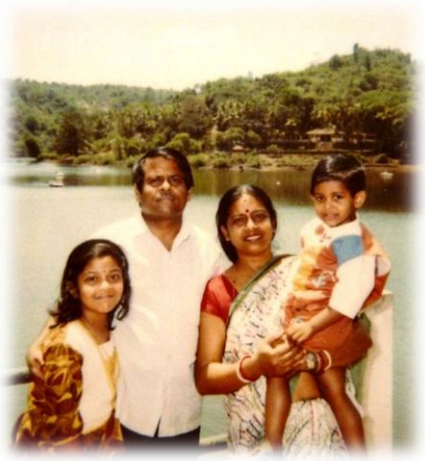
S. Banerji, A. Sinha, and C. Liu. Novel Color HWML Descriptors for Scene and Object Image Classification. In *The 3rd International Conference on Image Processing Theory, Tools and Applications*, Istanbul, Turkey, October 15-18, 2012.

S. Banerji, A. Sinha, and C. Liu. Scene Image Classification: Some Novel Descriptors. In *IEEE International Conference on Systems, Man, and Cybernetics*, Seoul, Korea, October 14-17, 2012.

S. Banerji, A. Sinha, and C. Liu. Novel Color, Shape and Texture-based Scene Image Descriptors. In *2012 IEEE International Conference on Intelligent Computer Communication and Processing*, Cluj-Napoca, Romania, August 30-September 1, 2012.

A. Sinha, S. Banerji, and C. Liu. Gabor-Based Novel Color Descriptors for Object and Scene Image Classification. In *The 16th International Conference on Image Processing, Computer Vision, and Pattern Recognition*, Las Vegas, Nevada, USA, July 16-19, 2012.

S. Banerji, A. Sinha, and C. Liu. Object and Scene Image Classification Using Unconventional Color Descriptors. In *The 16th International Conference on Image Processing, Computer Vision, and Pattern Recognition*, Las Vegas, Nevada, USA, July 16-19, 2012.



To My Beloved Parents and My Dearest Younger Brother

ACKNOWLEDGMENT

Upon completion of my doctoral dissertation, I would like to express my heartfelt appreciation to those people, whose guidance, advice and support have made my study and research more enriching and enjoyable in the Department of Computer Science at NJIT. First, I would take this opportunity to thank my dissertation advisor, Dr. Chengjun Liu, for his invaluable advice, sound guidance and his confidence in me. Dr. Liu has always provided immense encouragement and constant support which helped me achieve research exposure and enabled me in publishing several papers in peer reviewed conferences and journals.

Second, I express my gratitude to Dr. James Geller, Dr. Durgamadhab Misra, Dr. David Nassimi and Dr. Vincent Oria for serving on my committee. I want to thank them for the time they have spent to provide me with their valued feedback and suggestions on my research. I would also like to thank my lab mates, Sugata Banerji, Shuo Chen and Qingfeng Liu, for their support and assistance. Specially, I would always be indebted to my friend Sugata Banerji, who has been a constant source of support and a valued mentor in the four years of my doctoral study. I must also thank Ms. Angel Butler and Dr. George Olsen in the Computer Science department who has helped me in academic issues that I have had during these years.

Third, but most importantly, I am grateful to my family to be on my side for every decision I have taken. I thank my parents, Mr. Tapan Kumar Sinha and Mrs. Malina Sinha, for having faith in me and allowing me to pursue PhD at NJIT. Without them, I could not have accomplished this goal. I am also thankful to my brother, Souvik Brata Sinha, who has always given me the affection and comfort needed during the tough times of PhD, and inspired me to be a role model for him.

Finally, I would like to thank those people who have stayed with me over these four years, enriching my graduate life, outside research. In particular, I thank my friends:

Ankur Agrawal, Kashif Qazi, Amrita Banerjee, Sumit Chakraborty, Sumana Pai, Kirtan Shah, Suchandra Roy, and last but not the least, Suryadip Chakraborty for bearing me in both the hard and happy days. Also, I would like to apologize to all those who touched my life and helped me in various ways during these four years, but whose names I failed to mention here. I will remain ever thankful to them.

TABLE OF CONTENTS

Chapter **Page**

TABLE OF CONTENTS

(Continued)

Chapter **Page**

TABLE OF CONTENTS

(Continued)

Chapter **Page**

LIST OF TABLES

Table **Page**

LIST OF TABLES
(Continued)

Table **Page**

LIST OF FIGURES

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

CHAPTER 1

INTRODUCTION

The area of content-based image classification, search and retrieval is a rapidly-expanding research area. Due to the easy availability of cheap data storage, inexpensive cameras, fast computing power and increasing data transfer rates, enormous volumes of images are uploaded and shared over the Internet these days, which necessitates the development of a framework that can classify images into different categories automatically, and also identify the contents on providing a query image to perform efficient search and retrieval.

Understanding the semantics and contents of images for recognition remains one of the most difficult and prevailing problems in the machine intelligence and computer vision community (Li et al. 2010). With high variations in pose, angles, illumination and occlusions, object and scene classification is a very challenging task (Li et al. 2010; Torralba et al. 2003; Murphy et al. 2003; Lazebnik et al. 2006, 2004). A key step towards building a good classification framework includes discriminatory feature extraction. To this end, this dissertation attempts to develop new image descriptors. As the human visual system is much more efficient and accurate than any machine-based image classification approach, a representation that is modeled on the human visual cortex is much more likely to be better than other image representations for classification tasks. This motivates to propose innovative image descriptors by incorporating cues from the human visual system. In addition, local, texture, shape as well as color information are also integrated to construct robust and effective feature representations that are suitable for content-based image classification.

The Gabor wavelets, whose kernels are similar to the 2D receptive field profiles of the mammalian cortical simple cells (Marcelja 1980) exhibit desirable characteristics of spatial locality and orientation selectivity (Liu and Wechsler 2002; Liu 2004b). As the Gabor wavelet representation captures the local information corresponding to spatial frequency (scale), spatial localization, and orientation selectivity, it encodes images in a

manner so as to facilitate object and scene image classification.

Shape and texture are other cues based on which human beings visually distinguish between object and scene categories, and hence they contribute significantly to object and scene image classification. A popular technique for describing the local object appearance and shape within an image is calculating the Histograms of Oriented Gradients (HOG) that stores distribution of edge orientations within an image (Bosch et al. 2007b; Ludwig et al. 2009). The Pyramid Histograms of Oriented Gradients (PHOG) descriptor is a well-known shape descriptor that is inspired from the HOG and the spatial pyramid representation proposed by (Lazebnik et al. 2006). Recent works employing local texture features such as Local Binary Patterns (LBP) (Ojala et al. 1994; Zhu et al. 2010; Crosier and Griffin 2008), for example, have shown promising results for recognition and classification of texture and scene images (Banerji et al. 2011). In texture recognition, a Gabor filter-based approach has been successfully used (Manjunath and Ma 1996). Some researchers have also employed the LBP histogram sequences of the Gabor wavelets for face image recognition (Lee et al. 2010; Zhang et al. 2005). Fusion of local patterns of Gabor magnitude and phase for face recognition was found to be effective in (Xie et al. 2010).

In addition, color also provides powerful discriminating information as humans can distinguish thousands of colors, compared to about only two dozen shades of gray (Gonzalez and Woods 2001), and color images have been shown to perform better than grayscale images for image classification tasks (Liu and Mago 2012; Banerji et al. 2011; Liu 2011; Liu and Yang 2009; Liu 2007, 2004a; Shih and Liu 2005; Verma et al. 2010). Notable contributions on color based image classification appear in (Verma et al. 2010; Liu 2008, 2006; Liu and Mago 2012) that propose several new color spaces and methods for face, object and scene image classification. Global color features such as the color histograms and local invariant features provide varying degrees of success against image variations such as rotation, viewpoint and lighting changes, clutter and occlusions (Burghouts and Geusebroek 2009). Some desirable properties of the descriptors defined in different color

spaces include relative stability over changes in photographic conditions such as varying illumination. It has been shown that fusion of color features achieve higher classification performance in the works of (Banerji et al. 2011; Verma and Liu 2011; Stokman and Gevers 2007).

This dissertation explores several novel image descriptors based on texture, shape, color and local features from an image. Specifically, first, a new color Gabor-HOG (GHOG) descriptor is introduced by concatenating the Histograms of Oriented Gradients (HOG) of the component images produced by applying Gabor filters in multiple scales and orientations to encode shape information. Second, the GHOG descriptor is analyzed in six different color spaces and grayscale to propose different color GHOG descriptors, which are further combined to present a new Fused Color GHOG (FC-GHOG) descriptor. Third, a novel Gabor-PHOG (GPHOG) descriptor is proposed which improves upon the Pyramid Histograms of Oriented Gradients (PHOG) descriptor, and subsequently a new FC-GPHOG descriptor is constructed by combining the multiple color GPHOG descriptors and employing the Principal Component Analysis (PCA). Next, the Gabor-LBP (GLBP) is derived by accumulating the Local Binary Patterns (LBP) histograms of the local Gabor filtered images to encode texture and local information of an image. Furthermore, a novel Gabor-LBP-PHOG (GLP) image descriptor is proposed which integrates the GLBP and the GPHOG descriptors as a feature set. The GLBP and the GHOG descriptors are then combined to produce the Gabor-LBP-HOG (GLH) feature vector which performs well on different object and scene image categories. The six color GLH vectors are further concatenated to form the Fused Color GLH (FC-GLH) descriptor. Finally, the Wigner based Local Binary Patterns (WLBP) descriptor is proposed that combines multi-neighborhood LBP, Pseudo-Wigner distribution of images and the popular bag of words model to effectively classify scene images.

The classification performance of the proposed image descriptors is evaluated using two frameworks. In the first one, the new image descriptors are subjected to dimensionality

reduction by PCA and feature extraction using Enhanced Fisher Model (EFM). Then a nearest neighbor classifier is used to test their performance on several widely used and publicly available image datasets. In the other method, a Support Vector Machine (SVM) classifier is used for reporting the performance. The descriptors are shown to achieve a better classification performance than other popular image descriptors, such as the Scale Invariant Feature Transform (SIFT), the Pyramid Histograms of visual Words (PHOW), the Pyramid Histograms of Oriented Gradients (PHOG), Spatial Envelope (SE), Color SIFT four Concentric Circles (C4CC), Object Bank (OB), the Context Aware Topic Model (CA-TM), the Hierarchical Matching Pursuit (HMP) as well as LBP and a few others.

This dissertation is organized in the following manner. Chapter 2 discusses the related work by other researchers that have been used in this dissertation. Chapter 3 discusses object and scene image classification by introducing the Gabor-HOG (GHOG) and the FC-GHOG descriptors. Chapter 4 presents the novel GPHOG descriptor and evaluates its classification performance which improves upon the popular PHOG descriptor. Chapter 5 proposes two novel descriptors, the GLP and the FC-GLP descriptor, that incorporate color, shape, texture and local information from an image. Chapter 6 presents the GLH and the FC-GLH feature vectors that outperform both GLBP and GHOG. Chapter 7 proposes a part based image representation method by utilizing the Pseudo-Wigner distribution of images, the LBP, DCT smoothing, bag of words model and spatial pyramid representation techniques. Chapters 3, 4, 5, 6 and 7 also show the results of experiments done on various image datasets. Chapter 4 also provides a more detailed discussion of the experimental results to further evaluate the performance of the GPHOG descriptor on different categories of various image datasets. Finally, Chapter 8 summarizes the contributions of this dissertation and discusses future directions for research.

CHAPTER 2

BACKGROUND

A digital image is stored as a matrix of values and can be represented by a two-dimensional function $f(x,y)$ defined over the spatial domain where the value of the function at some particular x and y gives the image intensity at that point. Each of these discrete intensity values of the matrix is known as a picture element, or "pixel". A color image is defined by a function of two spatial variables and one spectral variable. Color images thus contain three such intensity matrices and can reproduce colors by storing three intensity values for each pixel of an image.

Color images contain more discriminative information than grayscale images and the color cue has been applied to facilitate image retrieval (Liu and Mago 2012; Liu 2011; Liu and Yang 2009; Liu 2006) and object, texture and scene search (Verma et al. 2010; Banerji et al. 2011). However, using the complete color information for feature extraction requires high computing power as well as more memory since color images contain at least three times the information contained in grayscale images. Discriminative information can be captured from color images by means of color features such as color invariants, color histogram and color texture. Some of the early methods for image classification were mainly based on the global descriptors such as the color and texture histograms (Niblack et al. 1993; Pontil and Verri 1998; Schiele and Crowley 2000). One such representative method is the color indexing system designed by Swain and Ballard, which used the color histogram for image retrieval from a large image database (Swain and Ballard 1991). More recently, the work of (Verma et al. 2010; Liu and Mago 2012; Liu 2008) on color based image classification propose several new color spaces and methods for face, object and scene classification. The HSV color space is used for scene category recognition in (Bosch et al. 2008), and the evaluation of local color invariant descriptors is performed in (Burghouts and Geusebroek 2009). The discriminating color space has been discussed in (Liu 2008).

Six popular color spaces and grayscale has been used in this dissertation for constructing discriminatory feature vectors, suitable for classification and search. These color spaces are discussed in detail in Section 2.1.

Efficient retrieval requires a robust feature extraction method that is able to extract meaningful low-dimensional patterns from very high dimensional data (Liu 2003). Low-dimensional representation is also important for achieving efficiency in computation. Principal Component analysis (PCA) has been a popular method for performing dimensionality reduction in image indexing and retrieval systems (Liu and Wechsler 2000). Section 2.2 discusses this technique. The Enhanced Fisher Model (EFM) feature extraction method has achieved good success for the task of image representation and retrieval (Liu and Wechsler 2000). This dissertation uses two classification frameworks. One technique performs EFM feature extraction followed by the Nearest Neighbor (NN) classification method for assigning class labels to test images. This combination, called the EFM-NN classifier henceforth, is explained in Section 2.3. The second classification framework uses a Support Vector Machine classifier (Vapnik 1995) which is discussed in Section 2.4.

2.1 Color Spaces

This section briefly reviews the six color spaces and grayscale used to define the proposed descriptors. Perception of color by the human visual system is made possible by specialized retinal cells called cone cells that contain pigments with different spectral sensitivities. The presence of three types of cones in the human eye sensitive to three different spectra results in trichromatic color vision. This is why, any system for representing the full visible color spectrum requires three variables which form a three-dimensional color space. Each color image, therefore, can be split up into three intensity images that are known as color component images or color planes.

The RGB color space, whose three component images represent the red, green, and blue primary colors, is the common tristimulus space for color image representation on a

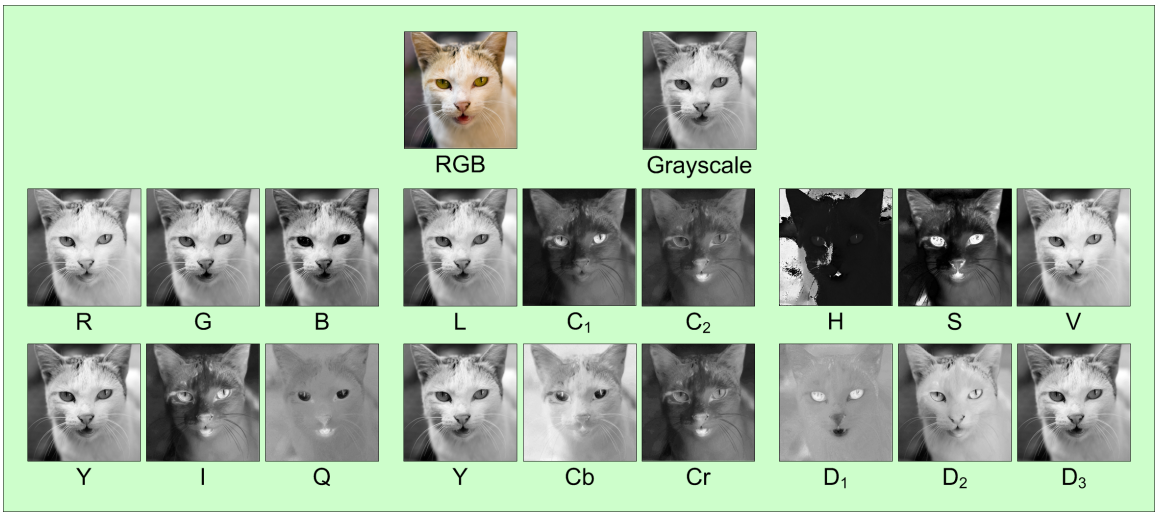


Figure 2.1 A color image, its grayscale image, and the color component images in the RGB, oRGB, HSV, YIQ, YCbCr and DCS color spaces, respectively.

computer. Other color spaces are usually derived from the RGB color space using either linear or nonlinear transformations.

The HSV (hue, saturation, and value) color space, however, is derived nonlinearly from the RGB color space (Smith 1978):

$$
\begin{aligned}
H &= \begin{cases} 60(\frac{G-B}{\delta}) & \text{if } MAX = R \\ 60(\frac{B-R}{\delta}+2) & \text{if } MAX = G \\ 60(\frac{R-G}{\delta}+4) & \text{if } MAX = B \end{cases} \\
S &= \begin{cases} \delta/MAX & \text{if } MAX \neq 0 \\ 0 & \text{if } MAX = 0 \end{cases} \\
V &= MAX
\end{aligned}
\tag{2.1}
$$

where $MAX = max(R,G,B)$, $MIN = min(R,G,B)$, and $\delta = MAX - MIN$.

The remaining four color spaces used in this dissertation are, again, transformed from the RGB color space using linear transformations.

The YCbCr color space is defined as follows (Gonzalez and Woods 2008):

$$\begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} + \begin{bmatrix} 65.481 & 128.553 & 24.966 \\ -37.797 & -74.203 & 112.000 \\ 112.000 & -93.786 & -18.214 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{2.2}$$

The YIQ color space is defined as given below (Shih and Liu 2005):

$$\begin{bmatrix} Y \\ I \\ Q \end{bmatrix} = \begin{bmatrix} 0.2990 & 0.5870 & 0.1140 \\ 0.5957 & -0.2745 & -0.3213 \\ 0.2115 & -0.5226 & 0.3111 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{2.3}$$

The three component images L, C_1, and C_2 of the oRGB color space are defined as follows (Bratkova et al. 2009):

$$\begin{bmatrix} L \\ C_1 \\ C_2 \end{bmatrix} = \begin{bmatrix} 0.2990 & 0.5870 & 0.1140 \\ 0.5000 & 0.5000 & -1.0000 \\ 0.8660 & -0.8660 & 0.0000 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{2.4}$$

The Discriminating Color Space (DCS) (Liu 2008) is derived from the RGB color space by means of discriminant analysis (Fukunaga 1990). The DCS defines discriminating component images via a linear transformation $W_D \in \mathbb{R}^{3\times 3}$ from the RGB color space

$$\begin{bmatrix} D_1 \\ D_2 \\ D_3 \end{bmatrix} = W_D \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{2.5}$$

where D_1, D_2, and D_3 are the values of the discriminating component images in the DCS color space. The transformation matrix $W_D \in \mathbb{R}^{3\times 3}$ may be derived through a procedure of discriminant analysis (Fukunaga 1990). Let S_w and S_b be the within-class and the between

class scatter matrices of the 3-D pattern vector $\mathscr{X}$, respectively where $S_w, S_b \in \mathbb{R}^{3x3}$. The discriminant analysis procedure derives a projection matrix W_D by maximizing the criterion $J_1 = tr(S_w^{-1} S_b)$ (Fukunaga 1990). This criterion is maximized when W_D^t consists of the eigenvectors of the matrix $S_w^{-1} S_b$ (Fukunaga 1990)

$$S_w^{-1} S_b W_D^t = W_D^t \Delta \tag{2.6}$$

where W_D^t, Δ are the eigenvector and eigenvalue matrices of $S_w^{-1} S_b$, respectively. Figure 2.1 shows a color image, its grayscale image, and its color component images in the RGB, oRGB, HSV, YIQ, YCbCr and DCS color spaces, respectively. The grayscale image here is an intensity image generated from the RGB image by forming a weighted sum of the R, G, and B components:

$$Gray = 0.2990R + 0.5870G + 0.1140B \tag{2.7}$$

Note that these are the same weights used to compute the Y component of the YIQ color space.

2.2 Principal Component Analysis (PCA)

Principal component analysis, or PCA, which is the optimal feature extraction method in the sense of the mean-square-error, derives the most expressive features for signal and image representation. Specifically, let $\mathscr{X} \in \mathbb{R}^N$ be a random vector whose covariance matrix is defined as follows (Fukunaga 1990):

$$S = \mathscr{E}\{[\mathscr{X} - \mathscr{E}(\mathscr{X})][\mathscr{X} - \mathscr{E}(\mathscr{X})]^t\} \tag{2.8}$$

where $\mathscr{E}(\cdot)$ represents expectation and t the transpose operation. The covariance matrix S is factorized as follows (Fukunaga 1990):

$$S = \Phi \Lambda \Phi^t \tag{2.9}$$

where $\Phi = [\phi_1 \phi_2 \cdots \phi_N]$ is an orthogonal eigenvector matrix and $\Lambda = diag\{\lambda_1, \lambda_2, \ldots, \lambda_N\}$, a diagonal eigenvalue matrix with diagonal elements in decreasing order.

Decorrelation is an important property of PCA, i.e. the components of the transformed data, $\mathscr{X}' = \Phi^t \mathscr{X}$, are decorrelated since the covariance matrix of $\mathscr{X}'$ is diagonal, $\Sigma_{X'} = \Lambda$, and the diagonal elements are the variances of the corresponding components. A second important property of PCA is its optimal signal reconstruction with respect to minimum Mean Square Error (MSE) when just a subset of the principal components is used to represent the original signal. A popular application of this second property is the extraction of the most expressive features of $\mathscr{X}$. Towards that end, a new vector $\mathscr{Y}$ is defined: $\mathscr{Y} = P^t \mathscr{X}$, where $P = [\phi_1 \phi_2 \cdots \phi_K]$, and $K < N$. The most expressive features of $\mathscr{X}$ thus define the new vector $\mathscr{Y} \in \mathbb{R}^K$, which consists of the most significant principal components.

2.3 The Enhanced Fisher Model and the Nearest Neighbor Classification Rule

Object and scene image classification using the new descriptors introduced in this dissertation is implemented using the Enhanced Fisher Model (EFM) for feature extraction (Liu and Wechsler 2000) and the Nearest Neighbor (NN) to the mean classification rule for classification. This EFM feature extraction and NN classification procedure is referred to as the EFM-NN classifier throughout this dissertation.

In pattern recognition, a popular method, Fisher's Linear Discriminant (FLD), applies first PCA for dimensionality reduction and then discriminant analysis for feature extraction. PCA is discussed in the previous section, and discriminant analysis often optimizes a criterion defined on the within-class and between-class scatter matrices S_w and S_b, which are defined as follows (Fukunaga 1990):

$$S_w = \sum_{i=1}^{L} P(\omega_i) \mathscr{E}\{(\mathscr{Y} - M_i)(\mathscr{Y} - M_i)^t | \omega_i\} \tag{2.10}$$

$$S_b = \sum_{i=1}^{L} P(\omega_i)(M_i - M)(M_i - M)^t \tag{2.11}$$

where $P(\omega_i)$ is *a priori* probability, ω_i represent the classes, and M_i and M are the means of the classes and the grand mean, respectively. One discriminant analysis criterion is J_1: $J_1 = tr(S_w^{-1} S_b)$, and J_1 is maximized when Ψ contains the eigenvectors of the matrix $S_w^{-1} S_b$ (Fukunaga 1990):

$$S_w^{-1} S_b \Psi = \Psi \Delta \tag{2.12}$$

where Ψ, Δ are the eigenvector and eigenvalue matrices of $S_w^{-1} S_b$, respectively. The discriminating features are defined by projecting the pattern vector $\mathscr{Y}$ onto the eigenvectors of Ψ:

$$\mathscr{Z} = \Psi^t \mathscr{Y} \tag{2.13}$$

$\mathscr{Z}$ thus contains the discriminating features for image classification.

The FLD method, however, often leads to overfitting when implemented in an inappropriate PCA space. To improve the generalization performance of the FLD method, a proper balance between two criteria should be maintained: the energy criterion for adequate image representation and the magnitude criterion for eliminating the small-valued trailing eigenvalues of the within-class scatter matrix (Liu and Wechsler 2000). As a result, the Enhanced Fisher Model (EFM) is developed to improve upon the generalization performance of the FLD method (Liu and Wechsler 2000). Specifically, the EFM method improves the generalization capability of the FLD method by decomposing the FLD procedure into a simultaneous diagonalization of the within-class and between-class scatter matrices (Liu and Wechsler 2000). The simultaneous diagonalization reveals that during whitening the eigenvalues of the within-class scatter matrix appear in the denominator. Since the small eigenvalues tend to encode noise (Liu and Wechsler 2000), they cause the whitening step to fit for misleading variations, and this leads to poor generalization performance. To enhance performance, the EFM method preserves a proper balance between the need that the

selected eigenvalues account for most of the spectral energy of the raw data (for representational adequacy), and the requirement that the eigenvalues of the within-class scatter matrix (in the reduced PCA space) are not too small (for better generalization performance) (Liu and Wechsler 2000).

2.4 Support Vector Machine

The Support Vector Machine (SVM) is a particular realization of statistical learning theory. The approach described by SVM, known as structural risk minimization, minimizes the risk functional in terms of both the empirical risk and the confidence interval (Vapnik 1995). SVM is built from two ideas: (i) a nonlinear mapping of the input space to a high-dimensional feature space, and (ii) designing the optimal hyperplane in terms of the maximal margin between the patterns of the two classes in the feature space. SVM is very popular and has been applied extensively for pattern classification, regression, and density estimation since it displays a good generalization performance.

Let $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_k, y_k), \mathbf{x}_i \in \mathbb{R}^N$, and $y_i \in \{+1, -1\}$ be k training samples in the input space, where y_i indicates the class membership of $\mathbf{x}_i$. Let ϕ be a nonlinear mapping between the input space and the feature space, $\phi : \mathbb{R}^N \rightarrow \mathscr{F}$, i.e., $\mathbf{x} \rightarrow \phi(\mathbf{x})$. The optimal hyperplane in the feature space is defined as follows:

$$w_0 \cdot \phi(\mathbf{x}) + b_0 = 0 \tag{2.14}$$

It can be proven (Vapnik 1995) that the weight vector w_0 is a linear combination of the support vectors, which are the vectors $\mathbf{x}_i$ that satisfy $y_i(w_0 \cdot \phi(\mathbf{x}_i) + b_0) = 1$:

$$w_0 = \sum_{support vectors} y_i \alpha_i \phi(\mathbf{x}_i) \tag{2.15}$$

where α_i's are determined by maximizing the following functional:

$$L(\alpha) = \sum_{i=1}^{k} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{k} \alpha_i \alpha_j y_i y_j \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \tag{2.16}$$

subject to the following constraints:

$$\sum_{i=1}^{k} \alpha_i y_i = 0, \alpha_i \geq 0, i = 1, 2, \ldots, k \tag{2.17}$$

From Eqs. 2.14 and 2.15, the linear decision function in the feature space can be derived

$$f(\mathbf{x}) = sign(\sum_{supportvectors} y_i \alpha_i \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}) + b_0) \tag{2.18}$$

It should be noted that the decision function (see Eq. 2.18) is defined by the dot products in the high dimensional feature space, where computation might be prohibitively expensive. SVM, however, manages to compute the dot products by means of a kernel function (Vapnik 1995)

$$K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \tag{2.19}$$

Three classes of kernel functions widely used in kernel classifiers, neural networks, and SVMs are polynomial kernels, Gaussian kernels, and sigmoid kernels (Vapnik 1995):

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j)^d, \tag{2.20}$$

$$K(\mathbf{x}_i, \mathbf{x}_j) = exp(\frac{-(\|\mathbf{x}_i - \mathbf{x}_j\|)^2}{2\sigma^2}), \tag{2.21}$$

$$K(\mathbf{x}_i, \mathbf{x}_j) = tanh(k(\mathbf{x}_i \cdot \mathbf{x}_j) + v), \tag{2.22}$$

where $d \in \mathbb{N}$, $\sigma > 0$, $k > 0$, and $v < 0$.

The SVM implementation used for the experiments presented in this dissertation is the one that is distributed with the VLFeat package (Vedaldi and Fulkerson 2010). The parameters of the support vector machine are tuned empirically using only the training data, and the parameters that yield the best average precision on the training data are used for classification of the test data. In particular, the cost parameter C has been empirically set to 1 for the best classification performance in the experiments described here.

CHAPTER 3

THE NOVEL COLOR GABOR-HOG (GHOG) IMAGE DESCRIPTORS

This chapter introduces a novel set of color image descriptors based on shape and Gabor wavelets for object and scene image classification. Specifically, first, a new Gabor-HOG (GHOG) descriptor is proposed for image feature extraction by concatenating the Histograms of Oriented Gradients (HOG) of the component images produced by applying Gabor filters in multiple scales and orientations to encode shape information of an image. Second, a comparative assessment of the classification performance of the GHOG descriptor is made in six different color spaces as well as in grayscale. Finally, a new Fused Color GHOG (FC-GHOG) descriptor is proposed for object and scene image classification by integrating the GHOG descriptors in the six different color spaces to further incorporate color information. Feature extraction for the proposed descriptors employ Principal Component Analysis (PCA) and Enhanced Fisher Model (EFM), and the nearest neighborhood is exploited for final classification. Experimental results using three benchmark datasets, the Caltech 256 object categories dataset, the MIT Scene dataset, and the UIUC Sports Event dataset show that the proposed new image descriptors achieve better image classification performance than other popular image descriptors, such as the Scale Invariant Feature Transform (SIFT), the Pyramid Histograms of Oriented Gradients (PHOG), Spatial Envelope (SE), the Color SIFT four Concentric Circles (C4CC), Object Bank (OB), Context Aware Topic Model (CA-TM), as well as LBP.

3.1 Gabor-Based New Image Descriptors

This section briefly reviews the Gabor wavelet representation, and then discusses the generation of the proposed new image descriptors based on Gabor wavelets, shape, color and local information for object and scene image classification. In particular, first, a new Gabor-

HOG (GHOG) descriptor is introduced for encoding both local and shape information of an image. Finally, a novel FC-GHOG descriptor is proposed that fuses the GHOG descriptors in the six different color spaces to further incorporate color information.

3.1.1 Gabor Wavelet Representation

This section discusses the Gabor wavelets and how images are represented by Gabor wavelet features. First a brief background of the Gabor wavelets is given. Then, a description of how the Gabor wavelet representations have been used in this research is made to derive new image descriptors for object and scene image classification.

The Gabor wavelet is considered to be a good model for human visual receptive fields and hence the Gabor wavelet-based approach has been successfully used in image analysis (Marcelja 1980; Jones and Palmer 1987; Daugman 1985, 1988). The kernels of the Gabor wavelets are similar to the 2-D receptive field profiles of the mammalian cortical simple cells (Marcelja 1980). The Gabor kernels or filters can be generated from one kernel or mother wavelet by dilation and rotation and hence they are all self-similar. Each kernel is a product of a Gaussian envelope and a complex plane wave. The Gabor wavelets exhibit desirable characteristics of spatial locality and orientation selectivity and capture the local structure corresponding to spatial frequency (scale), spatial localization, and orientation selectivity. The 2-D Gabor wavelets were first introduced by Daugman (Daugman 1993) for human iris recognition. Promising results in face recognition have been achieved using the Gabor wavelets by (Liu 2004b; Liu and Wechsler 2001, 2003, 2002; Xie et al. 2010; Lee et al. 2010; Zhang et al. 2005).

A Gabor filter (kernel, wavelet) is obtained by modulating a sinusoid with a Gaussian distribution. In a 2-D scenario such as images, a Gabor filter is defined as:

$$\mathscr{G}_{\nu,\theta,\alpha,\sigma,\gamma}(x,y) = \exp(-\frac{x'^2+\gamma^2 y'^2}{2\sigma^2})\exp(i(2\pi\nu x'+\alpha)) \tag{3.1}$$

where $x' = x\cos\theta + y\sin\theta$, $y' = -x\sin\theta + y\cos\theta$, and ν, θ, α, σ, γ denote the spatial frequency of the sinusoidal factor, orientation of the normal to the parallel stripes of a Gabor function, phase offset, standard deviation of the Gaussian kernel and the spatial aspect ratio specifying the ellipticity of the support of the Gabor function, respectively. The Gabor wavelet representation of an image is obtained by the convolution of the image with a family of Gabor kernels as defined by Eq. (3.1). The convolution of an image I and a Gabor kernel $\mathscr{G}$ is defined as:

$$O(x,y) = I(x,y) * \mathscr{G}(x,y) \tag{3.2}$$

The response $O(x,y)$ to each Gabor kernel is a complex function with a real part $\mathrm{Re}\{O(x,y)\}$ and an imaginary part $\mathrm{Im}\{O(x,y)\}$ which is expressed as:

$$O(x,y) = \mathrm{Re}\{O(x,y)\} + i\,\mathrm{Im}\{O(x,y)\}$$

The magnitude response $\|O(x,y)\|$ is as follows:

$$\|O(x,y)\| = \sqrt{\mathrm{Re}\{O(x,y)\}^2 + \mathrm{Im}\{O(x,y)\}^2} \tag{3.3}$$

In this work, even symmetric Gabor filters (Jain et al. 2000; Barbu 2009) have been used. For an even symmetric Gabor filter, the Eq. (3.1) can be reduced to have the following general form in the spatial domain:

$$\mathscr{G}_{\nu,\theta,\sigma,\gamma}(x,y) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right)\cos(2\pi\nu x') \tag{3.4}$$

where $x' = x\cos\theta + y\sin\theta$, $y' = -x\sin\theta + y\cos\theta$, and ν, θ, σ, γ denote the spatial frequency of the sinusoidal factor, orientation of the normal to the parallel stripes of a Gabor function, standard deviation of the Gaussian kernel and the spatial aspect ratio specifying the ellipticity of the support of the Gabor function, respectively. In this work, the magnitude responses of the Gabor wavelet representations have been used for subsequent construction of the novel descriptors. Going forward, the phrase Gabor filtered images is used to refer to the magnitude responses of the Gabor wavelet representations of images. For all the experiments in this chapter, the chosen parameter values are $\sigma = 8$, $\gamma = 1$, ν = [1/8, 1/16], and $\theta = [0, \pi/6, \pi/3, \pi/2, 2\pi/3, 5\pi/6]$.

3.1.2 The Gabor-HOG (GHOG) Descriptor

The Gabor wavelet representation captures local structure corresponding to spatial frequency (scale), spatial localization, and orientation selectivity (Schiele and Crowley 2000; Liu and Wechsler 2002) and hence multi-resolution and multi-orientation Gabor filtering has been used for extraction of the new feature vector. To further encode local and shape information from the Gabor feature representations, the novel Gabor-HOG (GHOG) descriptor is introduced by extracting the HOG features from the set of Gabor filtered images.

The Histograms of Oriented Gradients (HOG) (Dalal and Triggs 2005) method is inspired from the Scale Invariant Feature Transform (SIFT) (Lowe 2004; Chen and Liu 2012) and was first applied for human detection (Dalal and Triggs 2005). The idea of HOG rests on the observation that local object appearance and shape can often be characterized well by the distribution of local intensity gradients or edge directions (Dalal and Triggs 2005). HOG features are derived based on a series of well-normalized local histograms of image gradient orientations in a dense grid (Dalal and Triggs 2005). In particular, the image window is first divided into small cells. For each cell, a local histogram of the gradient directions or the edge orientations is accumulated over the pixels of the cell. All the histograms within a block of cells are then normalized to reduce the effect of illumination

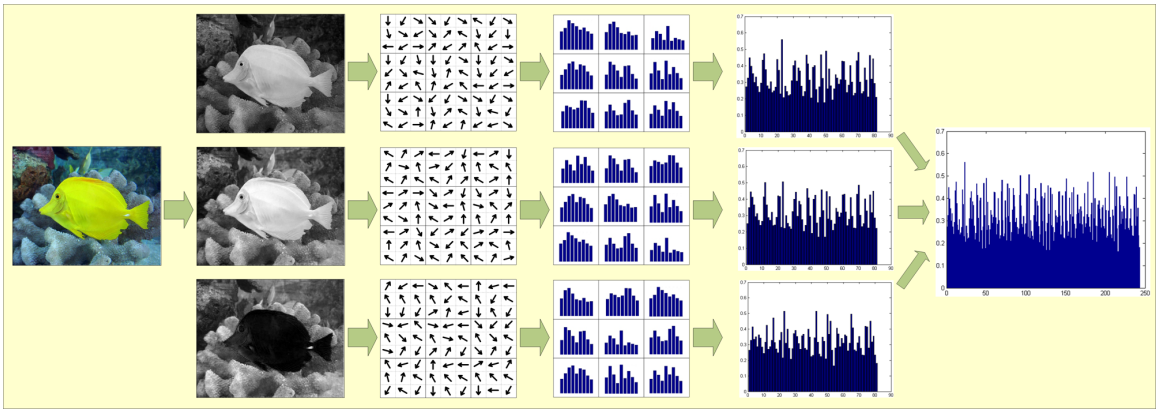


Figure 3.1 A color image, its three color component images, the orientation gradients of the different cells from every color component image, the histograms of orientation gradients formed from each of the cells of the color component images, the three HOG descriptors for the three color component images, and the concatenated HOG descriptor for the whole color image.

variations. The blocks can be overlapped with each other for performance improvement. The final HOG features are formed by concatenating all the normalized histograms into a single vector. For a color image, this process is repeated separately for the three component images and then the histograms are concatenated to form the color HOG descriptor. Figure 3.1 shows how the HOG descriptor is formed by the gradient histograms from a color image. Specifically, it shows a color image in the leftmost column. The three images in the second column are the three color component images of the original color image. The third column displays the gradient orientations of the color component images in the second column. Note that in this work, 3×3 cells have been used for deriving the orientation gradients from the images. Each of the three images in the fourth column shows the histograms of the orientation gradients for each cell. In this example, each of the histograms contains nine bins. The fifth column shows the HOG descriptors for the three color component images formed by concatenating the histograms of oriented gradients of the small cells, and finally the rightmost image shows the generation of the color HOG descriptor produced by concatenating the three HOG features corresponding to the three color component images.

The idea of the new Gabor-HOG (GHOG) descriptor is mainly based on the Gabor

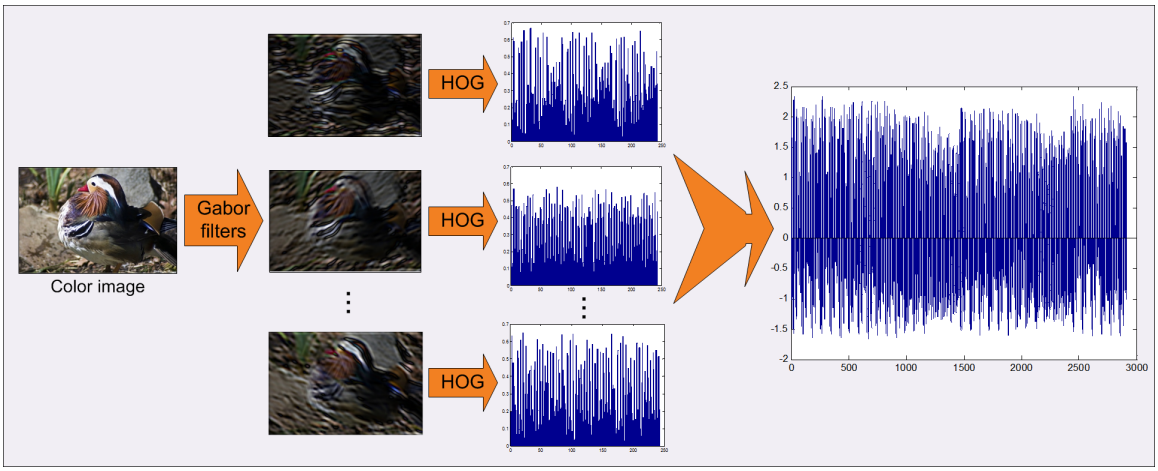


Figure 3.2 A color image, its Gabor filtered color images, the HOG descriptors obtained from the Gabor filtered color images, and the new GHOG descriptor derived from the concatenation and subsequent normalization of the color HOG descriptors of the Gabor filtered color images.

wavelet representations of an image and the HOG features. Specifically, first the Gabor filters defined in Eq. (3.4) are applied in two scales and six different orientations as discussed previously to encode local information corresponding to spatial frequency (scale), spatial localization, and orientation selectivity. For a color image, the Gabor filters are applied to the three color component images. Then the HOG features are derived from each color component of the Gabor filtered images that are obtained as a result of applying twelve different Gabor filters to the color image.

To compute the HOG features from the Gabor filtered images, 3×3 cells are used for deriving the orientation gradients from the images, where the histograms for each cell contain nine bins. The HOG features of each of the color components of all the Gabor-filtered images are then concatenated and finally normalized to zero mean and unit standard deviation to generate the novel GHOG descriptor. Figure 3.2 shows a color image, its Gabor filtered color images, the HOG descriptors obtained from the Gabor filtered color images, and the new GHOG descriptor derived by normalizing the concatenation of the color HOG descriptors of the Gabor filtered color images.

Figure 3.3 Some sample images from the Caltech 256 dataset.

3.1.3 The Fused Color GHOG (FC-GHOG) Image Descriptor

The color cue is often applied by the human visual system for recognizing images. Indeed, color provides powerful distinguishing information for pattern recognition in general and for object and scene image classification in particular (Liu and Mago 2012; Banerji et al. 2011; Liu 2011; Verma et al. 2010; Liu and Yang 2009; Liu 2008, 2007, 2006, 2004a). The motivation of the next descriptor is to further incorporate color information. In this section, an innovative Fused Color GHOG (FC-GHOG) descriptor is presented that fuses the most expressive features of the GHOG descriptors in six different color spaces, namely the RGB, oRGB, HSV, YIQ, DCS, and YCbCr color spaces (Liu 2008). The most expressive features of the GHOG descriptors are extracted by means of Principal Component Analysis (PCA).

The proposed FC-GHOG descriptor is derived by first computing the GHOG descriptors in the six different color spaces. Then the most expressive features from the six color GHOG descriptors are extracted using PCA and concatenated to create the novel FC-GHOG image descriptor.

3.2 Classifier Used

The proposed new descriptors presented in the preceding section are tested for classification by applying the Enhanced Fisher Model (EFM) for feature extraction (Liu and Wechsler

Figure 3.4 Some sample images from the MIT Scene dataset.

2000) and implementing the Nearest Neighbor (NN) to the mean classification rule for classification. This EFM-NN classifier has been described in detail in Section 2.3.

3.3 Experiments

This section briefly describes the datasets used and then reveals the experimental results.

3.3.1 Datasets

This section briefly discusses the three fairly challenging and popular datasets, namely: the Caltech 256 dataset, the MIT Scene dataset and the UIUC Sports Event dataset, on which the proposed descriptors are tested for performance evaluation.

Figure 3.5 Some sample images from the UIUC Sports Event dataset.

The Caltech 256 Dataset: The Caltech 256 dataset (Griffin et al. 2007) holds 30,607 images divided into 256 object categories and a clutter class. Each category contains a minimum of 80 images and a maximum of 827 images. The mean number of images per category is 119. The images represent a diverse set of lighting conditions, poses, backgrounds, and sizes (Griffin et al. 2007) and have high intra-class variability and high object location variability (Griffin et al. 2007). Most of the images are in color, in JPEG format with only a small percentage in grayscale. The average size of each image is 351×351 pixels. Some sample images from this dataset are shown in Figure 3.3.

The MIT Scene Dataset: The MIT Scene dataset (Oliva and Torralba 2001) (also known as the OT Scenes) has 2,688 images classified as eight scene categories: 360 coast, 328 forest, 260 highway, 308 inside of cities, 374 mountain, 410 open country, 292 streets, and 356 tall buildings. All of the images are in color, in JPEG format, and the size of each image is 256×256 pixels (Oliva and Torralba 2001). There is a large variation in light, content and angles, along with a high intra-class variation (Oliva and Torralba 2001). Figure 3.4 displays some sample images from this dataset.

The UIUC Sports Event Dataset: The UIUC Sports Event dataset (Li and Fei-Fei 2007) contains eight sports event categories: badminton (200 images), bocce (137 images), croquet (236 images), polo (182 images), rock climbing (194 images), rowing (250 images), sailing (190 images), and snowboarding (190 images). The mean image size is 845×1077 pixels. Most of the images are color jpeg images, with a small percentage in grayscale. A few sample images from this dataset are shown in Figure 3.5.

3.3.2 Comparison of the GHOG Descriptor in Different Color Spaces

In this section, a comparative assessment of the GHOG descriptor is made in six different color spaces – RGB, HSV, oRGB, YCbCr, DCS, and YIQ color spaces, as well as in grayscale, using the three datasets described earlier to evaluate classification performance. Towards that end, first the GHOG descriptor is derived from each image in the different

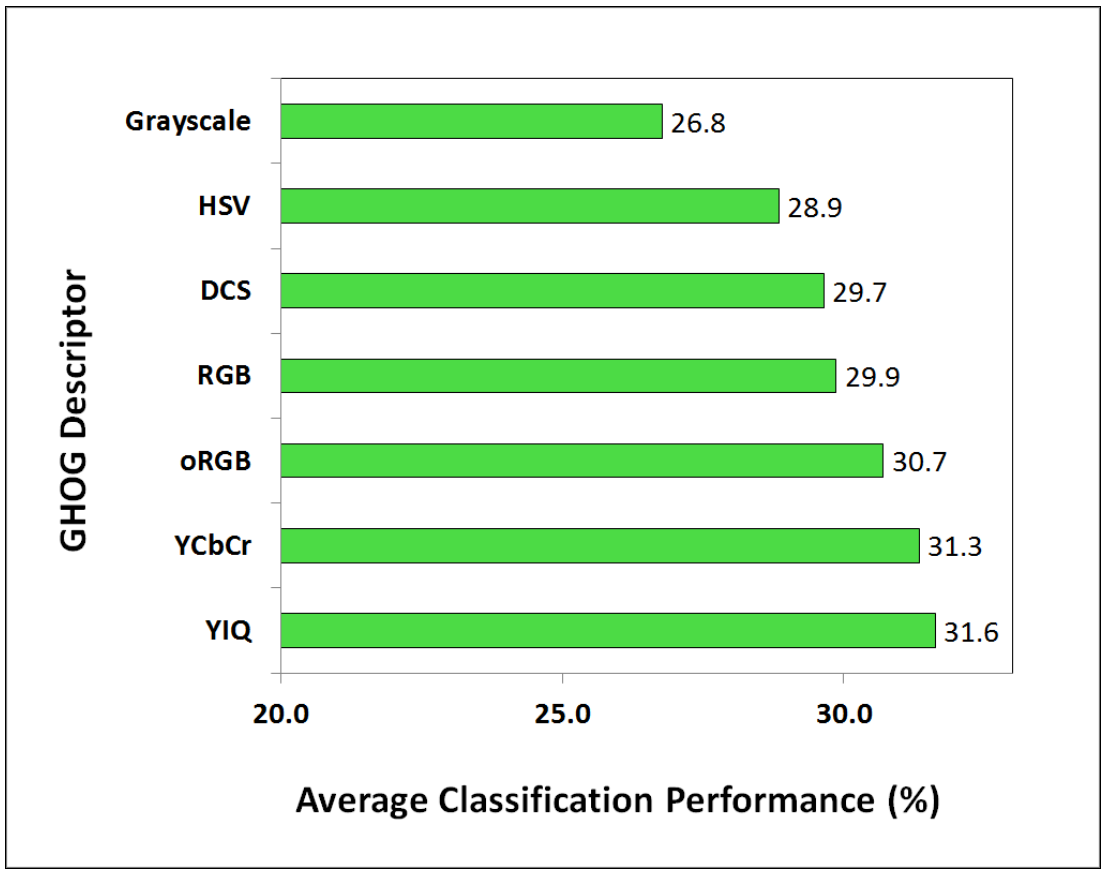


Figure 3.6 The average classification performance of the proposed GHOG descriptor in the YIQ, YCbCr, oRGB, RGB, DCS and HSV color spaces and also in grayscale using the EFM-NN classifier on the Caltech 256 dataset.

color spaces. Note that some large-scale images are resized in such a way that their largest dimension does not exceed 400 pixels. Each input image is converted into grayscale as well as transformed into images in the six color spaces. Each image in a single color space first undergoes Gabor filtering in six orientations and two scales to produce twelve different Gabor-filtered images. The HOG descriptor is further computed from these Gabor filtered images and concatenated which are normalized to zero mean and unit standard deviation to finally derive the GHOG feature, respectively. The EFM is applied for feature extraction and the nearest neighbor rule is finally used for image classification, where similarity score between a train and test vector is computed using the cosine similarity measure.

On the Caltech 256 dataset, experiments are conducted using a protocol defined

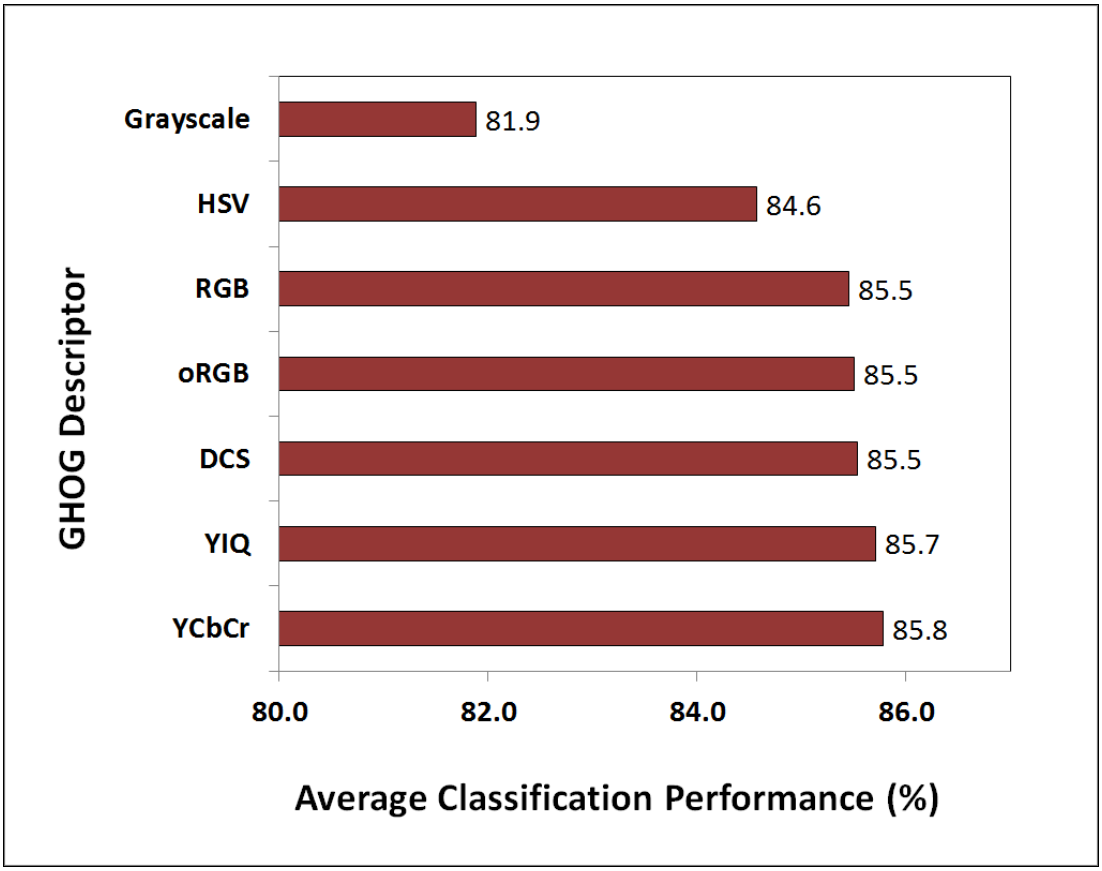


Figure 3.7 The average classification performance of the proposed GHOG descriptor in the YCbCr, YIQ, DCS, oRGB, RGB and HSV color spaces and also in grayscale using the EFM-NN classifier on the MIT Scene dataset.

in (Griffin et al. 2007). For each class, 50 images are used for training and 25 images for testing, and five runs of experiments are performed using the data splits that are provided on the Caltech website (Griffin et al. 2007). Figure 3.6 reveals the comparative classification performance of the proposed GHOG descriptor in six different color spaces and also in grayscale. The horizontal axis indicates the average classification performance, which is the percentage of correctly classified images averaged across the 256 classes and the five runs of the experiments, and the vertical axis shows the seven different GHOG descriptors in the six color spaces and grayscale. It shows that GHOG descriptor in YIQ color space performs best with 31.6% classification performance, followed by the GHOG descriptors in YCbCr, oRGB, RGB, DCS, HSV and grayscale with 31.3%, 30.7%, 29.9%, 29.7%, 28.9%

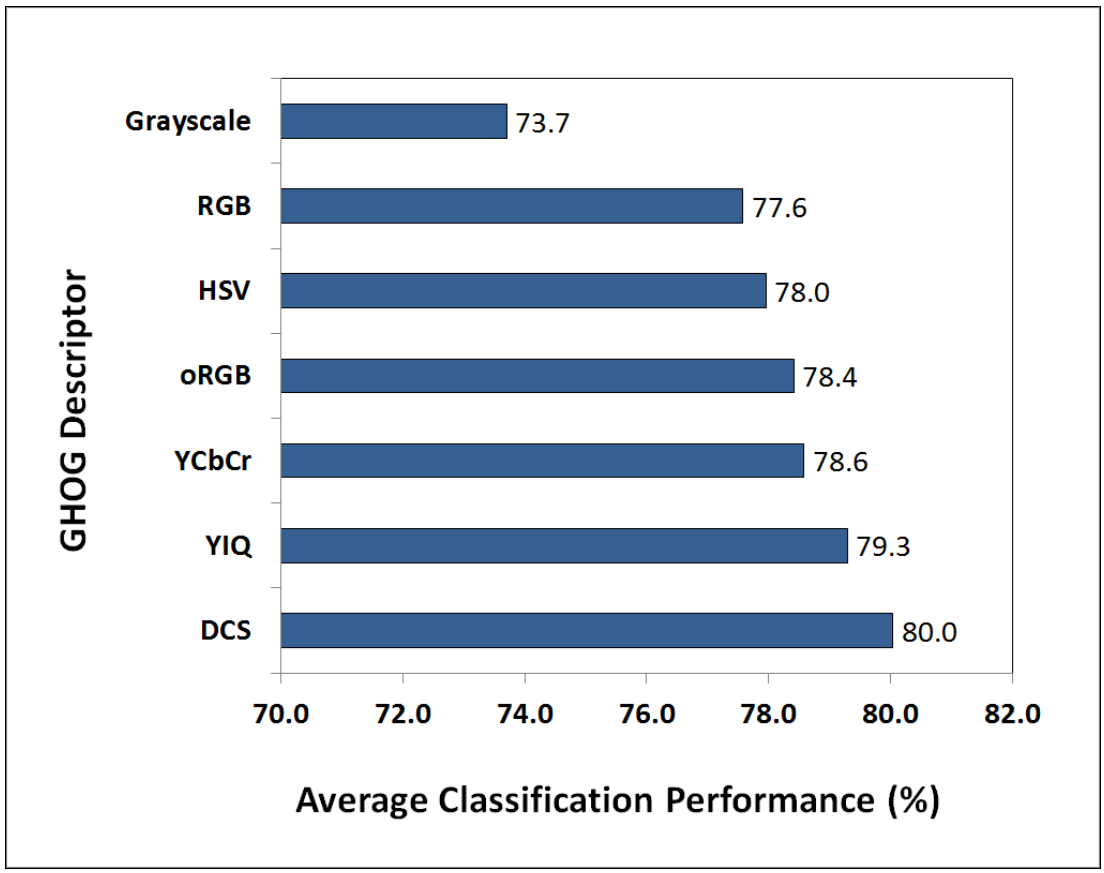


Figure 3.8 The average classification performance of the proposed GHOG descriptor in the DCS, YIQ, YCbCr, oRGB, HSV and RGB color spaces and also in grayscale using the EFM-NN classifier on the UIUC Sports Event dataset.

and 26.8% classification performances, respectively.

For the MIT Scene dataset, 100 images are used from each class for training and the rest of the images for testing. All experiments are performed for five random splits of the data to achieve more reliability. Figure 3.7 shows the detailed classification performance of the GHOG descriptor in six different color spaces and also in grayscale using the EFM-NN classifier. Here, the GHOG descriptor in YCbCr color space performs best with 85.8% classification performance, followed by the GHOG descriptors in YIQ, DCS, oRGB, RGB, HSV and grayscale with 85.7%, 85.5%, 85.5%, 85.5%, 84.6%, and 81.9% classification rates, respectively.

For the UIUC Sports Event dataset, a protocol defined in (Li and Fei-Fei 2007) is

used, which specifies that for each class in this dataset, 70 images are used for training and 60 images for testing. To achieve more reliable performance, the experiments are repeated five times using random splits of the data, and no overlapping occurs between the training and the testing sets of the same split. Figure 3.8 shows the detailed classification performance of the GHOG descriptor in grayscale and in six different color spaces using the EFM-NN classifier. Here also, the horizontal and vertical axes show the average classification performance and different descriptors in the six color spaces and in grayscale, respectively. It can be seen from this figure that the GHOG descriptor in DCS color space performs best with 80.0% classification performance, followed by the GHOG descriptors in YIQ, YCbCr, oRGB, HSV, RGB and grayscale with 79.3%, 78.6%, 78.4%, 78.0%, 77.6%, and 73.7% classification rates, respectively.

3.3.3 Comparison of the FC-GHOG Descriptor and Some Other Methods

Now, the performance of the proposed FC-GHOG descriptor is evaluated in the three datasets, and also compared with some popular descriptors. In particular, the FC-GHOG descriptor is first compared with the popular and robust SIFT-based Pyramid Histograms of visual Words (PHOW) descriptor (Bosch et al. 2007a) on all three datasets. For fair comparison, both descriptors apply the EFM-NN classifier for image classification. Then, the classification performance achieved by the FC-GHOG descriptor coupled with the EFM-NN classifier is compared to the image classification performance of some other state-of-the-art descriptors and classification approaches as reported in published papers.

To make a comparative assessment of the FC-GHOG descriptor with a popular SIFT-based descriptor, the Pyramid Histograms of visual Words (PHOW) feature vector (Bosch et al. 2007a) is generated using the software package VLFeat (Vedaldi and Fulkerson 2010). Here feature extraction is a three-step process. First, SIFT features are extracted from images using a fast SIFT process. In this algorithm, SIFT descriptors are computed at points on a dense regular grid instead of the SIFT-generated interest points (Lazebnik

Table 3.1 Comparison of the Classification Performance (%) of the FC-GHOG Descriptor with Other Popular Methods on the Caltech 256 Dataset

Descriptor		Performance (%)
oRGB-SIFT	(Verma et al. 2010)	23.9
gray-PHOW		25.9
color-PHOW		29.9
CSF	(Verma et al. 2010)	30.1
FC-GHOG	**(Proposed)**	**34.7**

et al. 2006; Bosch et al. 2007a). Second, the SIFT features are subjected to K-means clustering with K=1000 to form a visual vocabulary. Finally, the images are spatially tiled into 2×2 parts and the histograms of visual words are computed for the SIFT features from each part. These four histograms are concatenated to generate the final PHOW feature vector. For a color image, the same process is repeated for the three color component images and the feature vectors are concatenated. To compare the classification performance of the proposed descriptor, gray PHOW as well as the color PHOW feature vectors are used.

Table 3.1 shows the comparison of the classification performance of the proposed FC-GHOG descriptor with that of other popular descriptors. In particular, on the Caltech 256 dataset, the FC-GHOG descriptor achieves the average classification performance of 34.7%, compared to the color-PHOW and the gray-PHOW descriptors with the average classification rates of 29.9% and 25.9%, respectively. It also outperforms the classification success achieved by oRGB-SIFT and Color Sift Fusion (CSF) descriptors which yield 23.9% and 30.1% classification rates, respectively. It should be noted here that all these descriptors apply the same EFM-NN classifier to achieve the classification performances as reported in Table 3.1.

On the MIT Scene dataset, FC-GHOG performance is evaluated by comparing it with both the gray and color PHOW features, and also to some other popular descriptor performances as reported in the published papers. Table 3.2 gives a detailed result of the classification performances obtained by the different descriptors on this dataset. Specif-

Table 3.2 Comparison of the Classification Performance (%) of the FC-GHOG Descriptor with Other Popular Methods on the MIT Scene Dataset

Descriptor		Performance (%)
CLF	(Banerji et al. 2011)	79.3
CGLF	(Banerji et al. 2011)	80.0
gray-PHOW		82.5
SE	(Oliva and Torralba 2001)	83.7
color-PHOW		84.3
CGLF+PHOG	(Banerji et al. 2011)	84.3
C4CC	(Bosch et al. 2006)	86.7
FC-GHOG	**(Proposed)**	**87.6**

ically, using 100 training images and rest of them for testing, the FC-GHOG correctly classifies 87.6% of the images, whereas the color and the gray PHOW descriptors achieve 84.3% and 82.5% classification success. It also lists a comparison of the classification performance of the FC-GHOG descriptor with some popular descriptors used by other researchers on the MIT Scene dataset. Please note that the classification results of the popular descriptors achieved by other researchers are reported directly from their published work. With 100 training images per class, the FC-GHOG descriptor again gives the best classification performance of 87.6%, as compared to Color SIFT four Concentric Circles (C4CC) (Bosch et al. 2006) with a classification performance of 86.7%, to CGLF+PHOG (Banerji et al. 2011) with a classification performance of 84.3%, to Spatial Envelope (SE) with a classification performance of 83.7%, to Color LBP Fusion (CLF) (Banerji et al. 2011) with a classification performance of 79.3%, and to Color Grayscale LBP Fusion (CGLF) (Banerji et al. 2011) with a classification performance of 80.0%.

Table 3.3 reports the comparison of the classification performance of the proposed FC-GHOG descriptor and other popular image descriptors on the UIUC Sports Event dataset. Here too, the FC-GHOG descriptor performs the best with 84.0% average classification accuracy, whereas the color and gray PHOW descriptors coupled with the EFM-NN classifier achieve 79.0% and 76.4% success rates, respectively. The FC-GHOG descriptor

Table 3.3 Comparison of the Classification Performance (%) of the FC-GHOG Descriptor with Other Popular Methods on the UIUC Sports Event Dataset

Descriptor		Performance (%)
SIFT+GGM	(Li and Fei-Fei 2007)	73.4
OB	(Li et al. 2010)	76.3
gray-PHOW		76.4
CA-TM	(Niu et al. 2012)	78.0
color-PHOW		79.0
FC-GHOG	**(Proposed)**	**84.0**

also performs better compared to the SIFT+GGM (Li and Fei-Fei 2007) method with classification performance of 73.4%, to Object Bank (OB) (Li et al. 2010) with classification performance of 76.3%, and to Context Aware Topic Model (CA-TM) (Niu et al. 2012) with classification performance of 78.0%.

3.4 Summary

The main contributions of this chapter are in the generation of novel image descriptors for object and scene image classification based on color, shape and Gabor wavelet transformation. In particular, a new Gabor-HOG (GHOG) descriptor is introduced to encode shape information of an image. Then a comparative assessment the classification performance of the new GHOG descriptor in six different color spaces – the RGB, the HSV, the YCbCr, the oRGB, the DCS and the YIQ – as well as in grayscale is made. Finally, a new FC-GHOG descriptor is presented for object and scene image classification by integrating the GHOG descriptors in the six different color spaces to further incorporate color information. Experimental results using three grand challenge datasets, the Caltech 256 object categories dataset, the MIT Scene dataset, and the UIUC Sports Event dataset show that the proposed new image descriptors achieve better image classification performance than other popular image descriptors, such as the Scale Invariant Feature Transform (SIFT), the Pyramid Histograms of Oriented Gradients (PHOG), Spatial Envelope (SE), the Color SIFT four

Concentric Circles (C4CC), Object Bank (OB), Context Aware Topic Model (CA-TM), as well as the Local Binary Patterns (LBP).

CHAPTER 4

THE NEW COLOR GABOR-PHOG (GPHOG) DESCRIPTORS

Chapter 3 discusses the HOG descriptor and introduces a Gabor-based new GHOG descriptor to encode the shape information of an image. The Pyramid Histograms of Oriented Gradients (PHOG) is a shape descriptor that is inspired from HOG and also captures spatial locality of an image. Gabor wavelets are known to selectively enhance high frequency local information in different orientations. This chapter presents a novel set of image descriptors that encode information from color, shape, spatial and local features of an image to improve upon the popular Pyramid of Histograms of Oriented Gradients (PHOG) descriptor for object and scene image classification. In particular, a new Gabor-PHOG (GPHOG) image descriptor created by enhancing the local features of an image using multiple Gabor filters is first introduced for feature extraction. A comparative assessment of the classification performance of the GPHOG descriptor is then made in grayscale and six different color spaces to further propose two novel color GPHOG descriptors that perform well on different object and scene image categories. An innovative Fused Color GPHOG (FC-GPHOG) descriptor is finally presented by integrating the Principal Component Analysis (PCA) features of the GPHOG descriptors in the six color spaces to combine color, shape and local feature information.

As the human visual system is much more efficient and accurate than any machine-based image classification approach, a representation that is modeled on the human visual cortex is likely to be better than other image representations for classification tasks. The Gabor wavelets, whose kernels are similar to the two-dimensional (2-D) receptive field profiles of the mammalian cortical simple cells (Marcelja 1980) exhibit desirable characteristics of spatial locality and orientation selectivity (Liu 2004b). The Gabor wavelets can be used in a variety of applications (Mao et al. 2012), and as it captures the local information corresponding to spatial frequency (scale), spatial localization, and orientation selectivity,

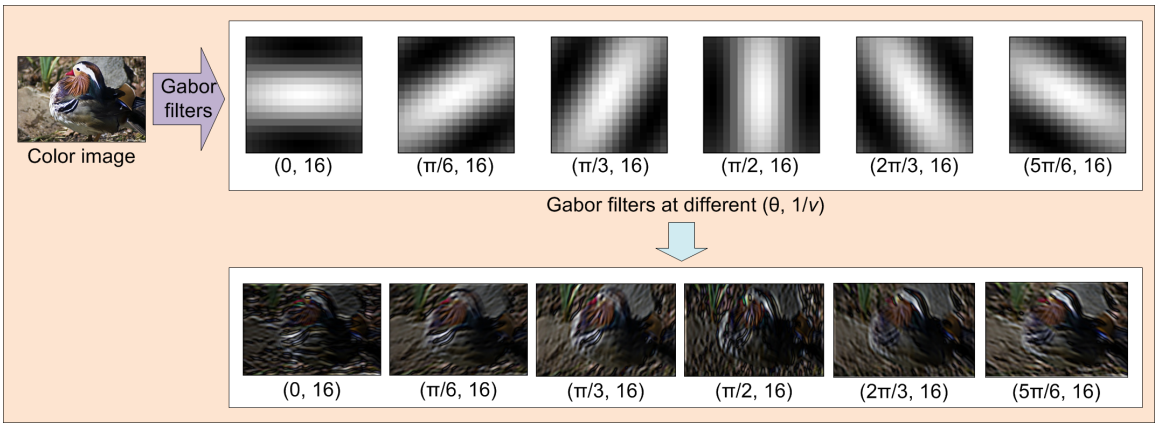


Figure 4.1 A color image, the Gabor filters (kernels) in one scale and six different orientations, and the magnitude responses of the Gabor wavelet representations of the color image on application of Gabor filters in different orientations. Each image labeled $(\theta, 1/\nu)$ corresponds to the Gabor-filtered image which is obtained by applying Gabor filter to the original color image with the specified orientation (θ) and scale (ν). Please note that the Gabor filters are enlarged for ease of display.

it encodes images in a manner so as to facilitate object and scene image classification. Gabor filtering has also been effectively used a pre-processing step before calculating other image descriptors such as the Local Binary Patterns (LBP) for face recognition (Lee et al. 2010).

The PHOG descriptor (Bosch et al. 2007b) represents local image shape and its spatial layout, together with a spatial pyramid kernel, by taking the histograms of the pixel gradients at different levels. Gabor wavelets, on the other hand, are known to capture local image information corresponding to spatial frequency (scale), spatial localization, and orientation selectivity. One of the motivations of this work is to improve the performance of the popular PHOG descriptor by preprocessing the image using a series of Gabor wavelet transformations, whose kernels are similar to the 2-D receptive field profiles of the mammalian cortical simple cells (Marcelja 1980).

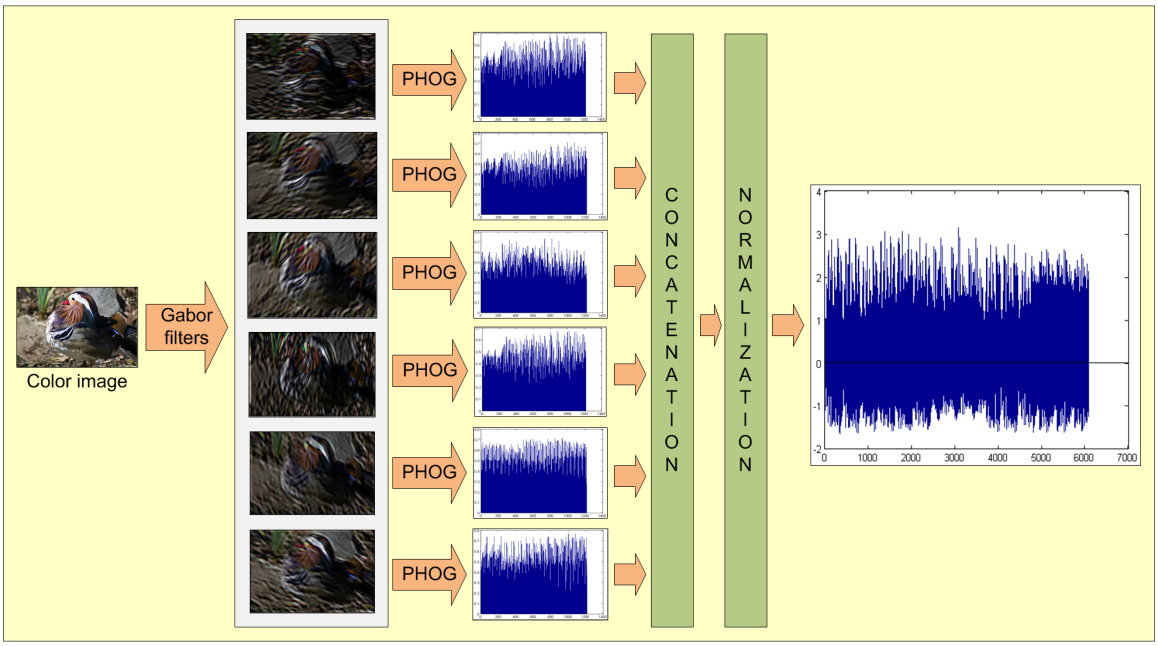


Figure 4.2 A color image, its Gabor filtered color images, the PHOG descriptors obtained from the Gabor filtered color images, and the new GPHOG descriptor derived from the concatenation and subsequent normalization of the color PHOG descriptors of the Gabor filtered color images.

4.1 New Image Descriptors Based on Color, Shape, and Wavelets

In this section, new image descriptors based on Gabor wavelets, color and shape are presented for object and scene image classification. In particular, first, a new Gabor-PHOG (GPHOG) descriptor is proposed for encoding local features, shape and the spatial layout of the shape within an image. Then a novel Fused Color GPHOG (FC-GPHOG) descriptor is constructed that integrates the PCA features of the GPHOG descriptors in six different color spaces to further incorporate color information.

4.1.1 The Gabor-PHOG (GPHOG) Descriptor

This section introduces a new Gabor-PHOG (GPHOG) descriptor that integrates the Gabor wavelet representation and the Pyramid of Histograms of Oriented Gradients (PHOG) to encode color, shape and local information from an image.

The Gabor wavelet is considered to be a good model for human visual receptive fields and hence the Gabor wavelet-based approach has been successfully used in image analysis (Marcelja 1980; Jones and Palmer 1987; Daugman 1988). The Gabor kernels or filters can be generated from one kernel or mother wavelet and it has been reviewed in Section 3.1.1. Even symmetric Gabor filters with the parameter values as $\sigma = 8$, $\gamma = 1$, ν = 1/16, and $\theta = [0, \pi/6, \pi/3, \pi/2, 2\pi/3, 5\pi/6]$ are used in this work. Figure 4.1 shows a color image, the Gabor filters (kernels) in one scale and six different orientations, and the magnitude responses of the Gabor wavelet representations of the original color image by applying six combinations of Gabor filters in multiple orientations as discussed above. The Gabor filtered images displayed in the figure are produced on application of Gabor filters with the specific (θ, $1/\nu$) parameter values as labeled in the figure. Please note that the size of Gabor filters used is 17 $\times$ 17, and those shown in the figure are enlarged for ease of display. It can be seen from the bottom row of Figure 4.1 that each of the magnitude responses of the six Gabor filters have the edges of one particular orientation enhanced. The orientation of the enhanced edges corresponds to the orientation of the Gabor filter that generated that response. This fact provides the motivation for the choice of the next step towards calculating the proposed feature vector.

To encode local features and shape information from the Gabor filtered images with enhanced edges, their Pyramid of Histograms of Oriented Gradients (PHOG) descriptors are computed. The PHOG (Bosch et al. 2007b) descriptor represents local image shape and its spatial layout and is inspired from the Histograms of Oriented Gradients (HOG) (Dalal and Triggs 2005) and the image pyramid representation of Lazebnik et al. (Lazebnik et al. 2006). Since PHOG encodes shape based on gradients around high frequency local features, it is logically a suitable descriptor to compute after the Gabor filtering operation. The presence of enhanced edges in the filtered images improves the information extracted by PHOG.

The PHOG descriptor adds spatial information to the HOG features following the

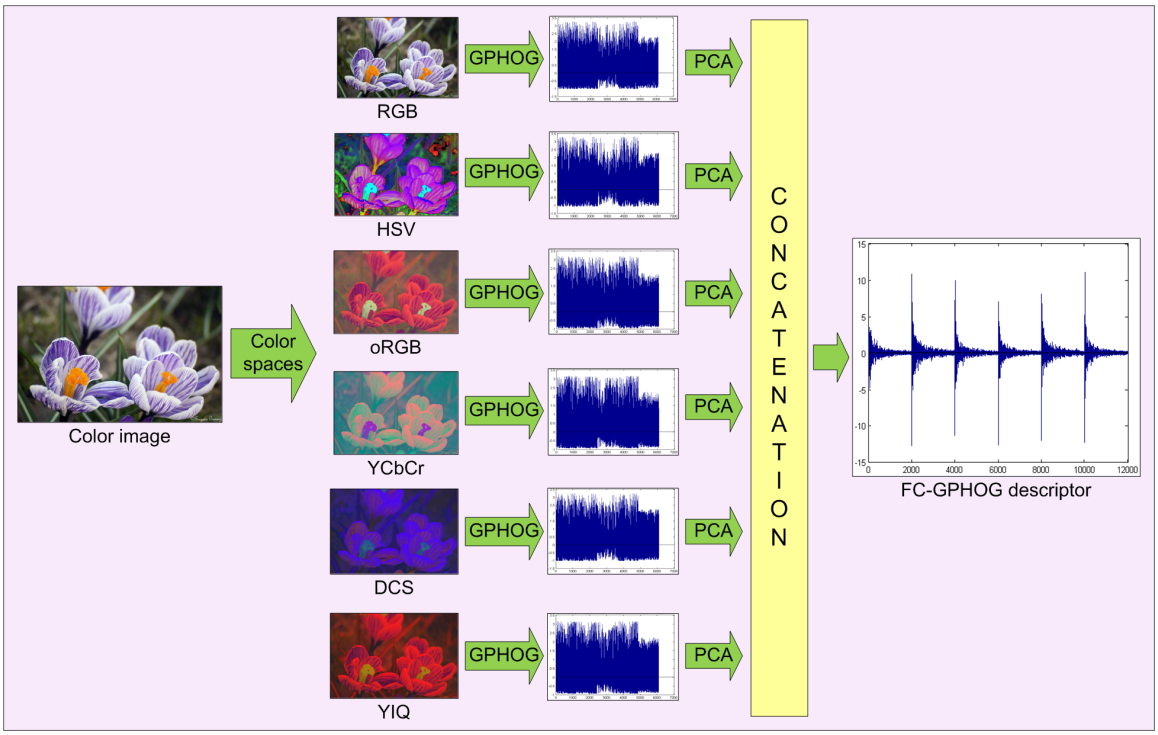


Figure 4.3 A color image, corresponding color images in the six color spaces, the GPHOG descriptors in the six color spaces, the PCA and the concatenation process, and the FC-GPHOG descriptor.

scheme based on spatial pyramid matching proposed by Lazebnik et al. (Lazebnik et al. 2006). To derive the PHOG features, each image is first divided into a sequence of increasingly finer spatial grids by repeatedly doubling the number of divisions in each axis direction (Bosch et al. 2007b). This is known as a pyramid representation as the number of points in a cell at one level is simply the sum over those contained in the four cells it is divided into at the next level (Bosch et al. 2007b). A HOG feature vector is computed for each grid cell at each pyramid resolution level. The final PHOG descriptor for the image is generated by concatenating all the HOG vectors obtained at each level. Thus, the PHOG encodes local shape by the distribution over edge orientations within a region, and the spatial layout is captured by tiling the image into regions at multiple resolutions. The distance between two PHOG image descriptors then reflects the extent to which the images contain similar shapes and correspond in their spatial layout (Bosch et al. 2007b). The PHOG

descriptor reduces to the HOG vector, which is a global edge or orientation histogram, if only the coarsest level is used for deriving the feature. It can enforce correspondence for tiles (spatial bins) over the image only if the finer levels are used. In this work, the PHOG descriptor is computed with only two levels. Again, for a color image, the entire process is repeated separately for the three component images and then the three PHOG features are concatenated to form the color PHOG descriptor.

The idea of the new GPHOG descriptor is based on the Gabor wavelet representations of an image and the PHOG features. The main motivation of the GPHOG is to enhance the PHOG features by applying a series of Gabor filters on the image before the computation of PHOG. Specifically, first the Gabor filters defined in Eq. (3.4) are applied in one scale and six different orientations as discussed previously to encode local information corresponding to spatial frequency (scale), spatial localization, and orientation selectivity. For a color image, the Gabor filters are applied to the three color component images. Then the PHOG features are derived from each color component of the Gabor filtered images obtained as a result of applying six different Gabor filters to the color image. To compute the PHOG features from the Gabor filtered images, a pyramid representation with two levels is used. The PHOG features of each of the color components of all the Gabor-filtered images are then concatenated and finally normalized to zero mean and unit standard deviation to generate the novel GPHOG descriptor. Figure 4.2 shows a color image, its Gabor filtered color images, the PHOG descriptors obtained from the Gabor filtered color images, and the new GPHOG descriptor derived by normalizing the concatenation of the color PHOG descriptors of the Gabor filtered color images.

4.1.2 An Innovative FC-GPHOG Descriptor

To make the GPHOG descriptor more suitable for image classification tasks, color information is further incorporated to generate the next descriptor. In this section, therefore an innovative Fused Color GPHOG (FC-GPHOG) descriptor is presented that fuses the most

expressive features of the GPHOG descriptors in six different color spaces, namely the RGB, oRGB, HSV, YIQ, YCbCr, and DCS color spaces (Liu 2008). The most expressive features of the GPHOG descriptors are extracted by means of Principal Component Analysis (PCA) (Fukunaga 1990). Figure 2.1 shows a color image, its grayscale image, and the color component images in the RGB, oRGB, HSV, YIQ, YCbCr, and DCS color spaces, respectively, which have been used in this paper.

To derive the proposed FC-GPHOG descriptor, the GPHOG descriptors are computed in the six different color spaces. The dimensionality of these six GPHOG descriptors are reduced using PCA, which derives the most expressive features in terms of minimum mean-square-error. Finally, the PCA features of the six color GPHOG descriptors are concatenated to create the novel FC-GPHOG image descriptor. Figure 4.3 shows a color image, its corresponding color images in the six color spaces, the GPHOG descriptors of the color images, the PCA process, the concatenation process, and the FC-GPHOG descriptor.

4.2 Classifier Used

After the descriptors have been generated, an efficient discriminatory feature extraction and classification method is required to achieve good classification accuracy. Here also, the Enhanced Fisher Model (EFM) for feature extraction is applied (Liu and Wechsler 2000) and the Nearest Neighbor (NN) rule is implemented for classification.

4.3 Experiments

The performance of the proposed descriptors for object and scene image classification is evaluated using two widely used and publicly available datasets – the MIT Scene dataset (Oliva and Torralba 2001), and the Caltech 256 dataset (Griffin et al. 2007). Three sets of experiments are done on these two datasets. The first set of experiments assesses the classification performance of the GPHOG descriptors in six different color spaces as well

as in grayscale.

The second set of experiments makes a comparison of the GPHOG and the traditional PHOG features for image classification and proves that the GPHOG descriptor is indeed better than PHOG for image classification. A five-fold cross validation of these results is made in six color spaces and grayscale. The PHOG descriptor from different color spaces is also fused and compared to the classification performance of the FC-GPHOG descriptor.

In the final set of experiments, the new FC-GPHOG descriptor is compared with other popular descriptors, such as the Scale Invariant Feature Transform (SIFT) (Lowe 1999) based Pyramid Histograms of visual Words (PHOW) (Bosch et al. 2007a) descriptor, the Color SIFT four Concentric Circles (C4CC) (Bosch et al. 2006), Spatial Envelope (Oliva and Torralba 2001), as well as Local Binary Patterns (LBP) (Ojala et al. 1994; Banerji et al. 2011).

Finally, the effect of different Gabor parameters on the classification performance and effectiveness of the proposed descriptors on different object and scene image categories of the two color image datasets are also discussed in detail.

4.3.1 Datasets

In this section, the two fairly challenging and popular datasets, namely: the MIT Scene dataset and the Caltech 256 dataset are described, on which the descriptors are tested for performance evaluation.

The MIT Scene Dataset: The MIT Scene dataset (Oliva and Torralba 2001) (also known as the OT Scenes) has 2,688 images classified as eight scene categories: coast, forest, highway, inside of cities, mountain, open country, streets, and tall buildings. A detailed description of this dataset is provided in Section 3.3.1. Figure 3.4 displays some sample images from this dataset.

The Caltech 256 Dataset: The Caltech 256 dataset (Griffin et al. 2007) holds

30,607 images divided into 256 object categories and a clutter class. Section 3.3.1 contains detailed description of this dataset. Figure 3.3 displays some sample images from this dataset.

4.3.2 Comparison of the GPHOG Descriptor in Different Color Spaces

In this section, a comparative assessment of the GPHOG descriptor is made in six different color spaces – RGB, HSV, oRGB, YCbCr, DCS and YIQ color spaces, as well as in grayscale, using the two datasets described earlier to evaluate classification performance. Towards that end, the GPHOG descriptor from each image is computed in the different color spaces. Note that for some large-scale images, they are resized in such a way that their largest dimension does not exceed 256 pixels. Each input image is converted into grayscale as well as transformed into the six color spaces. Each image in a single color space first undergoes Gabor filtering in six orientations and a single scale to produce six different Gabor-filtered images. The PHOG descriptors are further computed from these Gabor filtered images and concatenated to finally derive the GPHOG features, which are normalized to zero mean and unit standard deviation. The GPHOG descriptors are derived in the same manner from the images in six different color spaces, and also in grayscale.

For the MIT Scene dataset, two sets of experiments are performed. In the first set of experiments, 250 images from each class are chosen for training and the rest for testing. The experiments are repeated using five random splits of data to achieve more reliability. Figure 4.4 shows the detailed classification performance of the GPHOG descriptors in grayscale and in six different color spaces using the EFM-NN classifier. Note that the horizontal axis denotes the average classification performance, which is the percentage of correctly classified images averaged across all the eight classes and the five runs of experiments, and the vertical axis shows the different GPHOG descriptors in the six different color spaces and in grayscale. The GPHOG descriptors in the YCbCr and the YIQ color spaces perform the best with an average classification rate of 87.3% among all the GPHOG

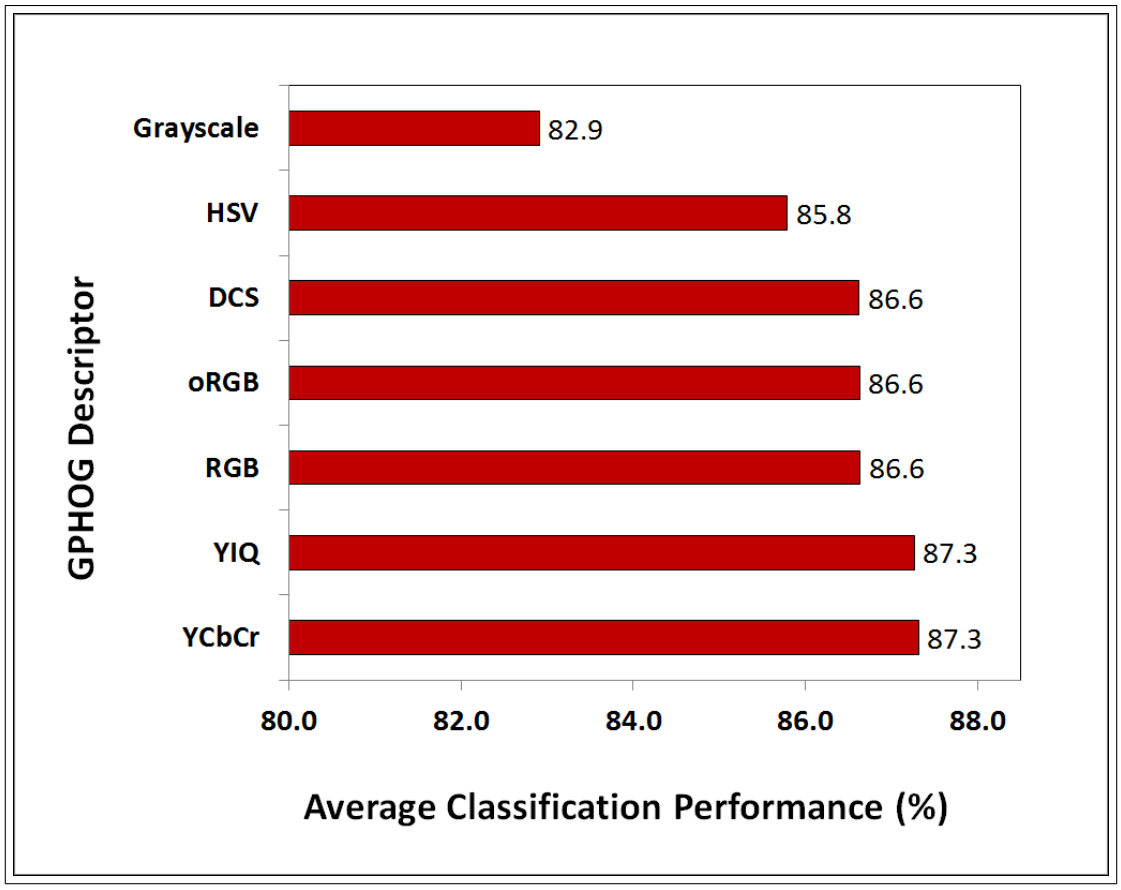


Figure 4.4 The average classification performance of the proposed GPHOG descriptor in the YCbCr, YIQ, RGB, oRGB, DCS, and HSV color spaces as well as in grayscale using the EFM-NN classifier on the MIT Scene dataset using 250 training images per class.

descriptors. Note that the GPHOG descriptor in grayscale performs the worst yielding an average success rate of 82.9% only, with a decrease of more than 4% from that achieved by the GPHOG in YCbCr and YIQ color spaces. This re-emphasizes the fact that adding color information is particularly suitable for classifying scene images. In the next set of experiments, 100 images from each class are used for training and the remaining images for testing. Figure 4.5 shows the classification performance of the GPHOG descriptors using this protocol and applying the EFM-NN classifier in the MIT Scene dataset. Among all the GPHOG descriptors, the GPHOG descriptors in the YIQ and the YCbCr color spaces outperform all the color as well as grayscale GPHOG descriptors and achieve a success rate of 84.0%.

On the Caltech 256 dataset, experiments are conducted using a protocol defined in (Griffin et al. 2007). For each class, 50 images are used for training and 25 images for testing, and five runs of experiments are done using the data splits that are provided on the Caltech website (Griffin et al. 2007). Figure 4.6 reveals the comparative classification performance of the proposed GPHOG descriptors in six different color spaces and also in grayscale. Here also, the GPHOG in the YIQ and YCbCr color spaces yield the best average classification performance of 30.1%.

It can be observed that for all the experiments that are conducted with the GPHOG descriptors in different color spaces on the two datasets, the GPHOG descriptors in the YIQ and YCbCr color spaces outperform the other color GPHOG descriptors, as revealed

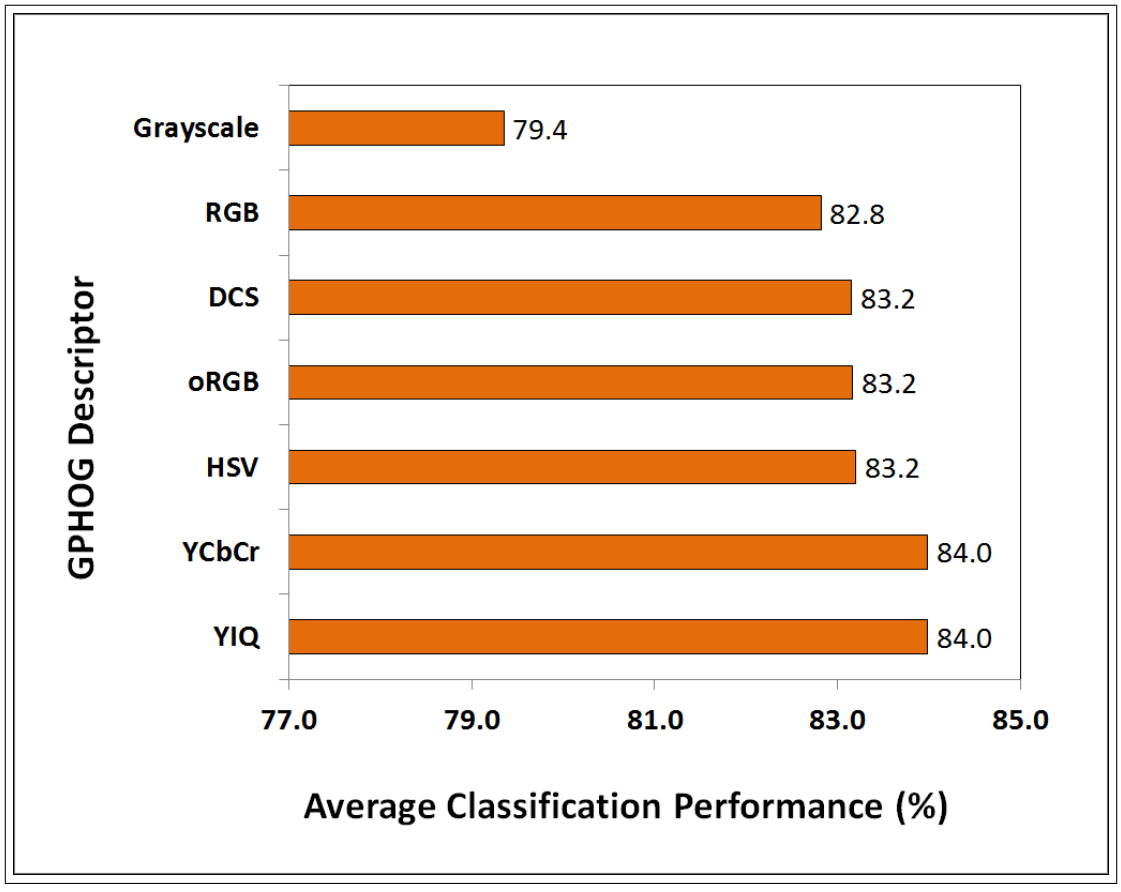


Figure 4.5 The average classification performance of the proposed GPHOG descriptor in the YIQ, YCbCr, HSV, oRGB, DCS, and RGB color spaces along with grayscale using the EFM-NN classifier on the MIT Scene dataset using 100 training images per class.

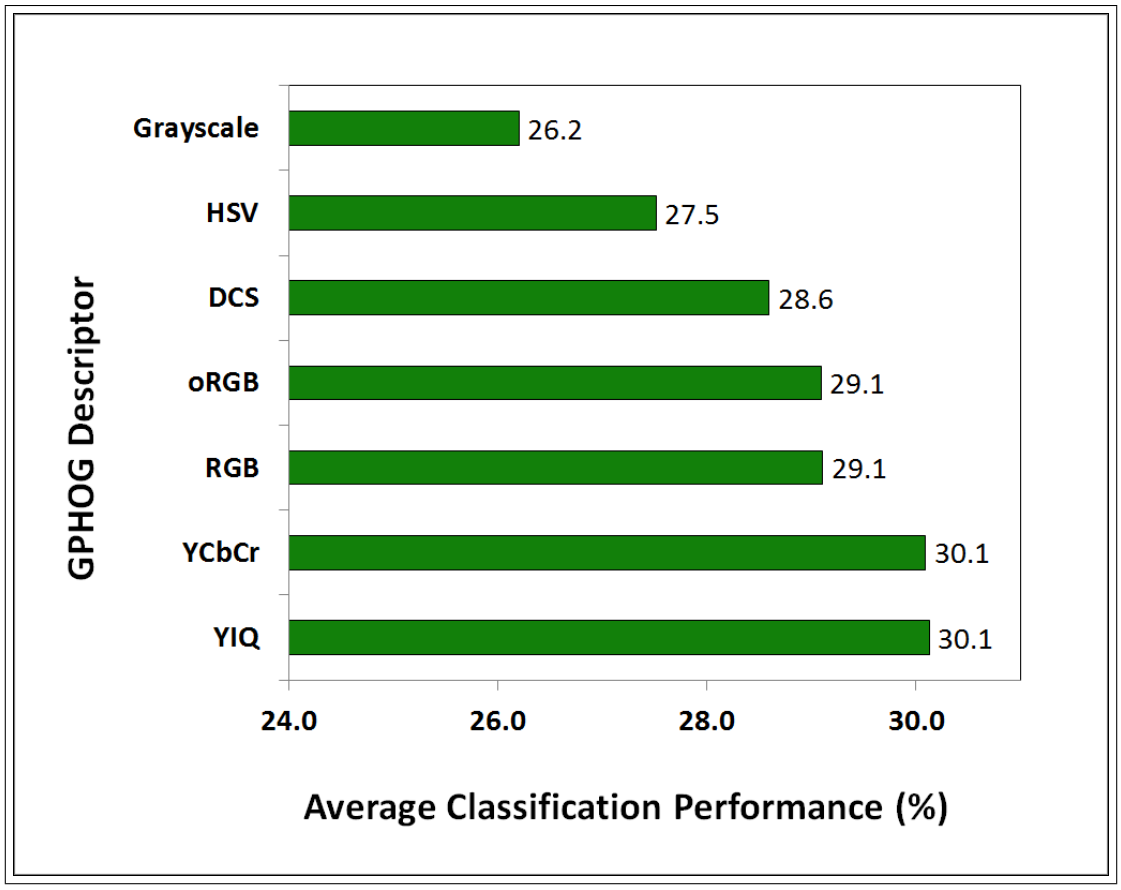


Figure 4.6 The average classification performance of the proposed GPHOG descriptor in the YIQ, YCbCr, RGB, oRGB, DCS, HSV color spaces and also in grayscale using the EFM-NN classifier on the Caltech 256 dataset.

by Figures 4.4, 4.5 and 4.6. Therefore two color GPHOG representations are proposed — the YIQ-GPHOG and the YCbCr-GPHOG — that are particularly more suitable for object and scene image classification among other color GPHOG features.

4.3.3 Comparison of the PHOG and GPHOG Descriptors

One of the primary motivations of this work is to improve the popular PHOG descriptor. In this section, the GPHOG descriptor is compared with the PHOG descriptor thereby establishing empirically the advantage of using the Gabor filtering step in the proposed framework, as opposed to not using it. Towards that end, the classification performance of

both the PHOG and GPHOG descriptors is assessed on the two datasets described earlier in six different color spaces as well as grayscale. Note that all the descriptors apply the same classification framework, i.e. the EFM-NN classifier. Further, the FC-GPHOG is also compared with the FC-PHOG features. Towards that end, first the FC-PHOG vector is created by combining the most expressive features of the six different color PHOG vectors, and then it is compared with the proposed FC-GPHOG.

Figure 4.7 displays the results of these experiments on the MIT Scene dataset, using 250 training samples from each class and five splits of experiments. It shows that the GPHOG clearly outperforms the PHOG in all the different color spaces and grayscale. Here the horizontal axis denotes the descriptors in different color spaces, while the vertical axis shows the mean average classification performance in percentage. This increase in classification rate is most significant in grayscale, where the GPHOG outperforms PHOG by more than 8%, followed by RGB where this difference is a little under 7%. Figure 4.8 shows similar results for the Caltech 256 dataset. Here the GPHOG again reveals a significant improvement over the PHOG descriptors in all the color spaces and grayscale. Here also, the most improvement occurs in the RGB color space and grayscale, with over 5% increase in the classification rates in both cases. These results clearly assert the importance of applying Gabor filtering before extracting PHOG for the construction of the proposed descriptors.

4.3.4 Comparison of FC-GPHOG with Other Popular Descriptors

The performance of the proposed FC-GPHOG descriptor is now evaluated on the two datasets, and also compared with some popular descriptors. In particular, first the FC-GPHOG descriptor is compared with the popular and robust SIFT-based Pyramid Histograms of visual Words (PHOW) descriptor (Bosch et al. 2007a). For fair comparison, both descriptors apply the EFM-NN classifier for image classification. Then the classification performance achieved by the FC-GPHOG descriptor is also compared with the

image classification performance of some other descriptors and classification approaches as reported in published papers.

To compare the proposed FC-GPHOG descriptor with the popular SIFT-based feature, the Pyramid Histograms of visual Words (PHOW) feature vector (Bosch et al. 2007a) is first generated using the software package VLFeat (Vedaldi and Fulkerson 2010). For both PHOW and FC-GPHOG, PCA is used to obtain the most expressive features and the EFM-NN classifier in order to make a fair comparison. Figure 4.9 reveals that the FC-GPHOG descriptor performs better than both the grayscale and the color PHOW descriptors on the two datasets. It also shows the average classification performance of the

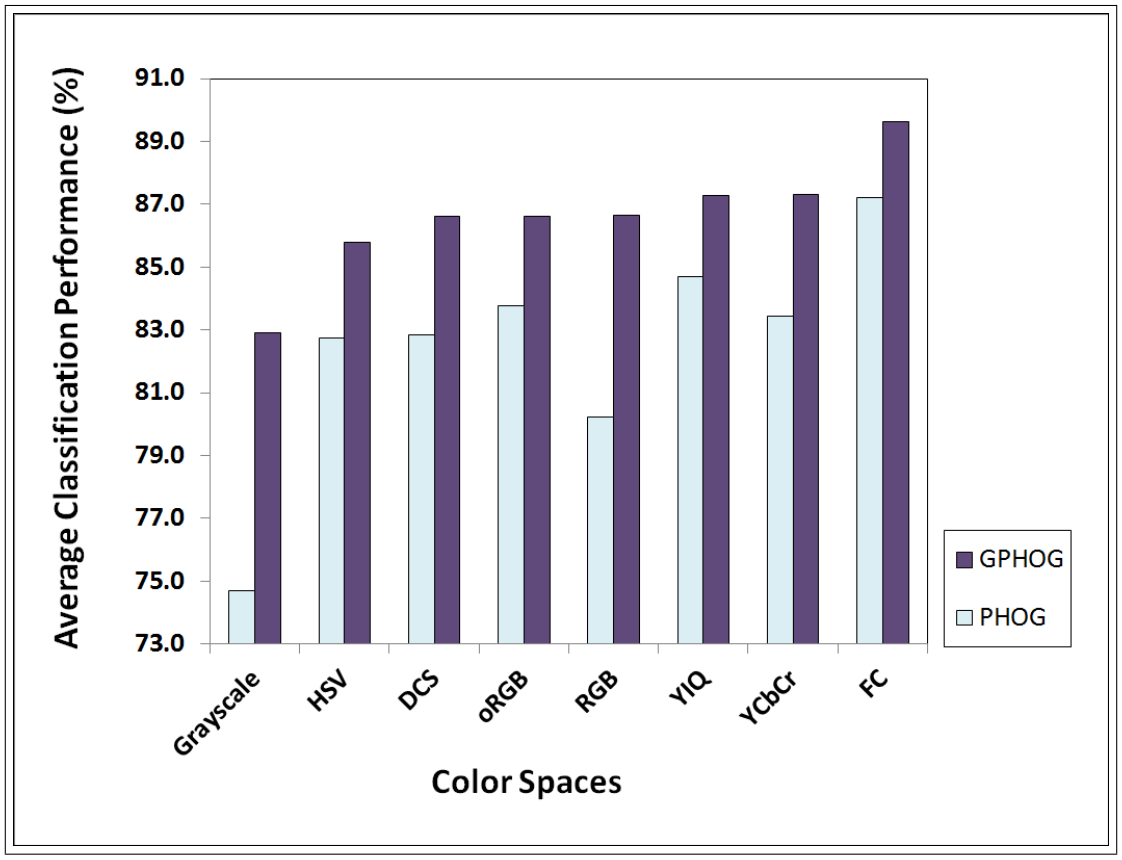


Figure 4.7 A comparison of the average classification performances of the PHOG and the proposed GPHOG descriptors in the grayscale, HSV, DCS, oRGB, RGB, YIQ and YCbCr color spaces, as well as the fusion of these color spaces on the MIT Scene (with 250 training images per class) dataset. Note that all these descriptors apply the EFM-NN classifier.

descriptors in the MIT Scene dataset using two protocols as explained in the preceding section.

A comparison of the classification performance of the FC-GPHOG descriptor is made with some popular descriptors used by other researchers on the MIT Scene dataset. Table 4.1 lists the classification success achieved by the proposed FC-GPHOG and some other descriptors on this dataset for the two sets of experiments. Please note that the classification results of the popular descriptors achieved by other researchers are reported directly from their published work. With 250 training images, the proposed FC-GPHOG descriptor achieves 89.6% classification accuracy which is at par with CGLF+PHOG (Banerji et al. 2011) and better than CGLF (Banerji et al. 2011). With 100 training images per class, the FC-GPHOG descriptor again gives the better classification performance of 86.0%, as compared to CGLF+PHOG (Banerji et al. 2011) and CGLF (Banerji et al. 2011).

4.3.5 Effect of Different Gabor Orientations on FC-GPHOG Descriptor

In this section, the impact of the different Gabor parameters on the classification performance of the FC-GPHOG descriptor is analyzed. The PHOG descriptor is first extracted in different color spaces from the six different Gabor filtered images formed as a result

Table 4.1 Comparison of the Classification Performance (%) of the FC-GPHOG Descriptor with Other Popular Methods on the MIT Scene Dataset

	#train = 2000, #test = 688	
FC-GPHOG	**Proposed Descriptor**	**89.6**
CGLF+PHOG	(Banerji et al. 2011)	89.5
CGLF	(Banerji et al. 2011)	86.6
	#train = 800, #test = 1888	
FC-GPHOG	**Proposed Descriptor**	**86.0**
C4CC	(Bosch et al. 2006)	86.7
CGLF+PHOG	(Banerji et al. 2011)	84.3
SE	(Oliva and Torralba 2001)	83.7
CGLF	(Banerji et al. 2011)	80.0

of using each of the six Gabor filters used in this work. Figure 4.10 illustrates the results obtained in detail. In particular, the horizontal axis shows the two datasets with the three experimental protocols as discussed earlier and the vertical axis shows the average classification performance. The results demonstrate that each of the orientations yield a classification performance that is significantly lower than the classification rate achieved by the FC-GPHOG descriptor as a whole (as shown in Figure 4.9). This indicates that each of the Gabor parameters encode non-redundant information and contribute towards the final result.

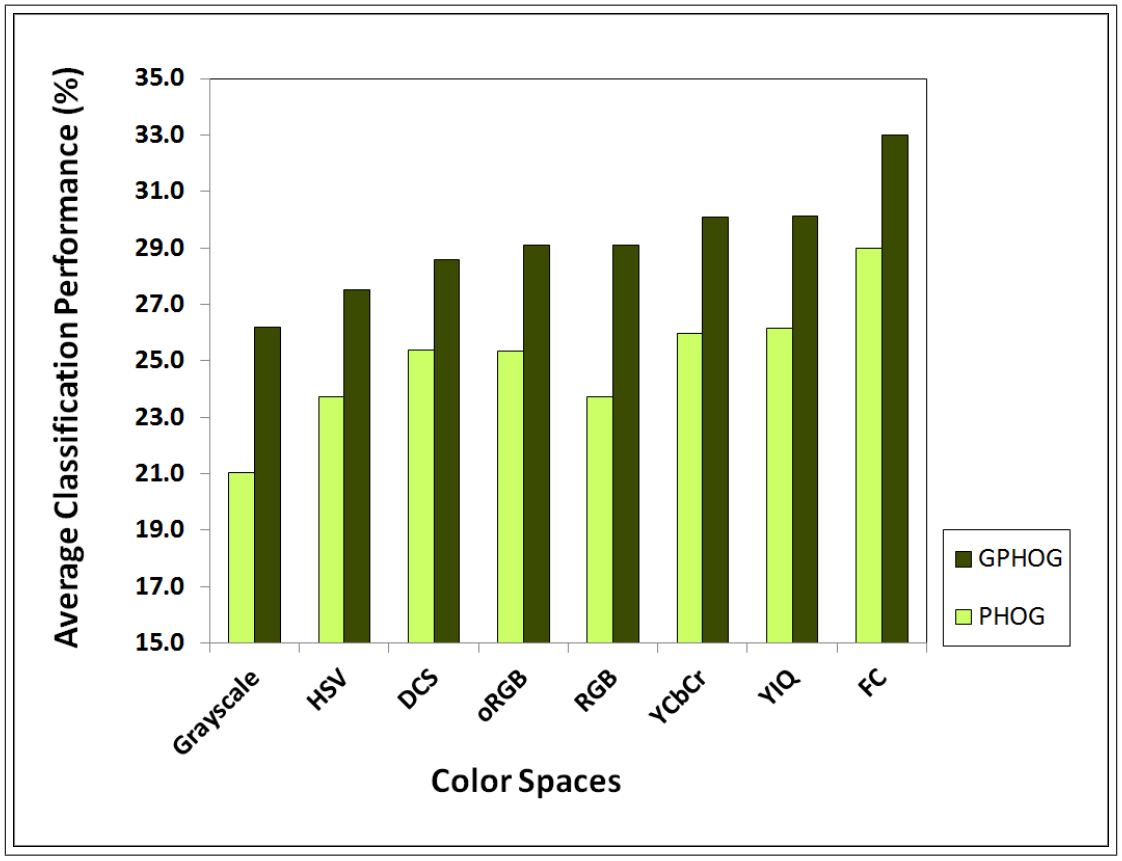


Figure 4.8 A comparison of the average classification performances of the PHOG and the proposed GPHOG descriptors in the grayscale, the HSV, the DCS, the oRGB, the RGB, the YCbCr and the YIQ color spaces, as well as the fusion of these color spaces on the Caltech 256 dataset. Note that all the descriptors apply the EFM-NN classifier.

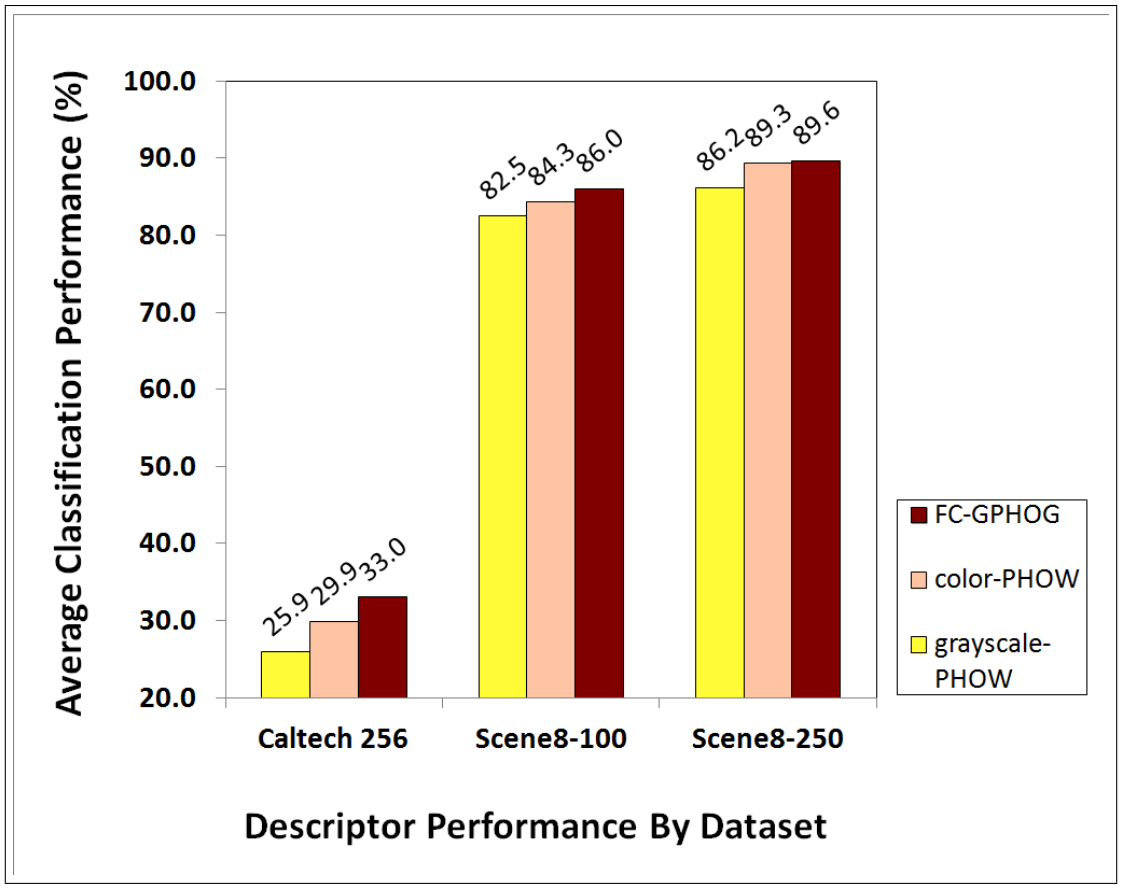


Figure 4.9 A comparison of the average classification performances of the grayscale-PHOW descriptor, the color-PHOW descriptor, and the proposed FC-GPHOG descriptor on the two image datasets – the Caltech 256, and the MIT Scene (with 100 and 250 training images per class) datasets. Note that all the three descriptors apply the EFM-NN classifier.

4.3.6 Class-wise Classification Performance of the GPHOG Descriptors

In this section, the class-wise classification success of the proposed descriptors are discussed on different object and scene image categories of the two color image datasets.

Figure 4.11 shows three classification confusion matrices between the eight categories in the MIT scene dataset with the categories in alphabetical order. In particular, the leftmost matrix represents classification based on the grayscale-GPHOG descriptor, the center matrix represents classification based on the YCbCr-GPHOG descriptor and the rightmost matrix represents classification based on the FC-GPHOG descriptor. In each confusion matrix, the rows show assigned classes while the columns show actual classes.

For instance, a high value at row 1, column 6 signifies that a lot of images from class 6 (open country) get assigned the class label 1 (coast). In the experiments, 250 images from each class were used for training.

It can be seen from Figure 4.11 that the best classified categories are 2 (forest), 8 (tall building), and 1 (coast) with success rates of 96.4%, 95.3%, and 93.1%, respectively. Category 6 (open country) is the most difficult category to classify. As the confusion matrix shows, some of the open country scenes are classified as coast, some as forest and some as mountain scenes. Parts (a), (b) and (c) of Figure 4.12 show some of the particularly confusing images from the open country category that get misclassified as coast, forest and mountain, respectively. The other three categories that are confused with each other are

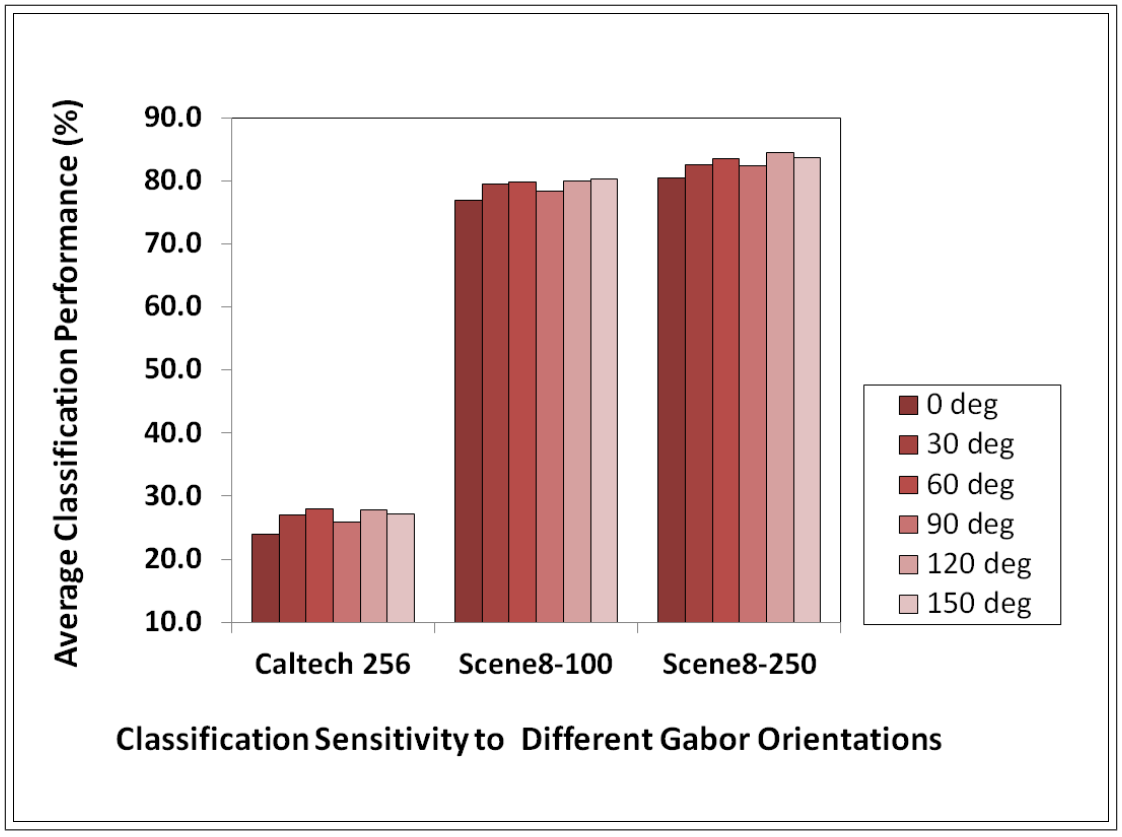


Figure 4.10 A comparison of the average classification performances achieved by using the different Gabor orientations of the FC-GPHOG descriptor on the two image datasets – the Caltech 256, and the MIT Scene (with 100 and 250 training images per class) datasets.

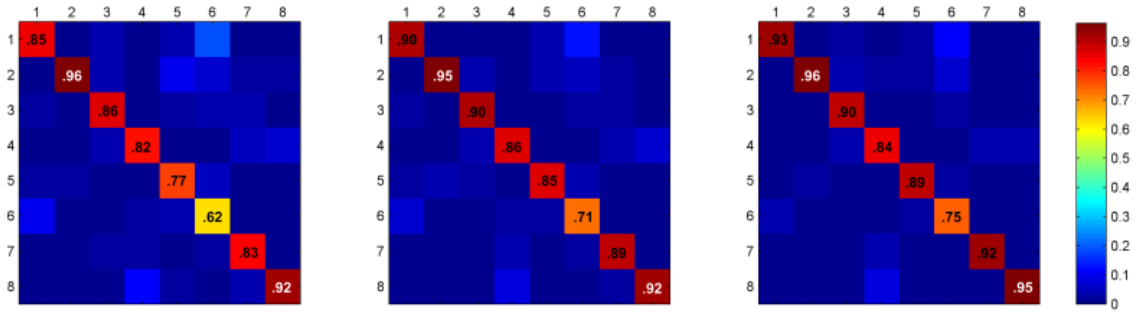


Figure 4.11 The confusion matrices between the eight categories in the MIT scene dataset with the categories in alphabetical order. The matrices from left to right represent classification using grayscale-GPHOG, YCbCr-GPHOG and FC-GPHOG, respectively. In each confusion matrix, the rows show assigned classes while the columns show actual classes.

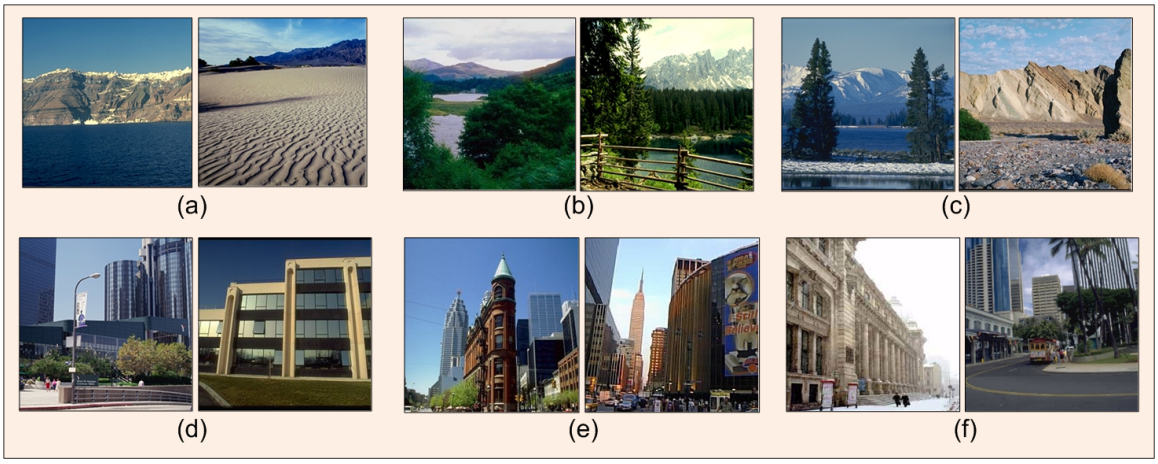


Figure 4.12 Some ambiguous images from the MIT scene dataset. Parts (a), (b) and (c) show some images from the open country category that get misclassified as coast, forest and mountain, respectively. Parts (d), (e) and (f) show ambiguous images from the inside city, tall building and street categories, respectively that contain similar features.

inside city, street and tall buildings. Parts (d), (e) and (f) of Figure 4.12 show two images each from the inside city, tall building and street categories, respectively that contain similar elements and hence cause misclassification. These results are similar to those reported by (Oliva and Torralba 2001) which would indicate that the confusion is due to an inherent ambiguity in the manual annotation of these particular dataset categories themselves.

Figure 4.13 displays the classification rates for all 256 categories of the Caltech

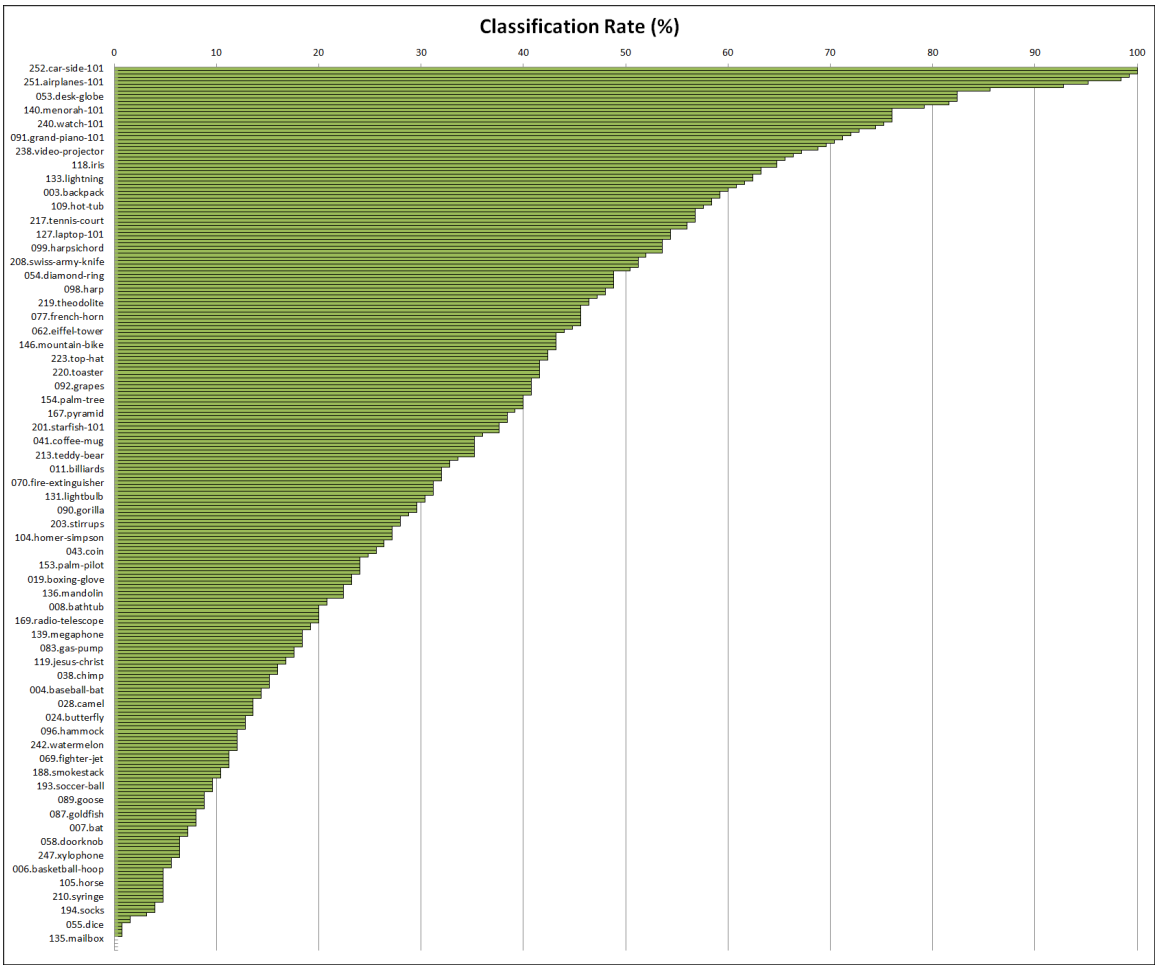


Figure 4.13 The average classification rates using the FC-GPHOG descriptor and EFM-NN classifier for all the categories of the Caltech 256 image dataset. Note that all category labels are not shown here to increase readability.

256 object categories dataset in descending order of the classification performance using the FC-GPHOG descriptor and EFM-NN classifier. In particular, it shows that the average classification performance varies widely among the different classes of this dataset, ranging from 0% to 100%. Please note that all categories have not been labeled in Figure 4.13 to increase readability. This dataset is much more complex and varied in its composition of categories and hence it is difficult to explain the classification performance on this dataset using one-to-one category misclassifications. For instance, Figure 4.14 shows some images from different classes that contain human figures, and it would be impossible to completely

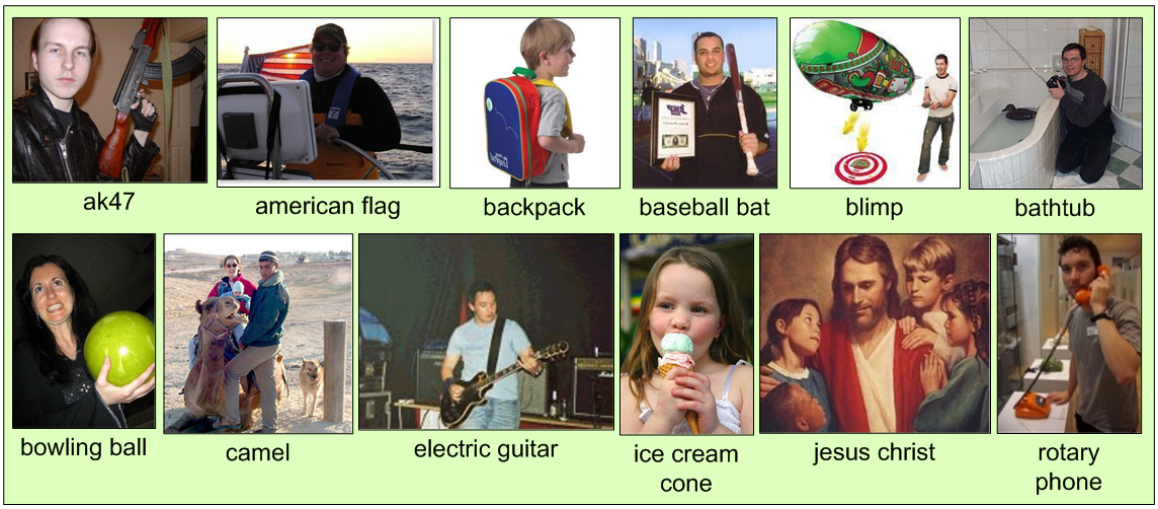


Figure 4.14 Some example images from the Caltech 256 object categories dataset. Note that none of these sample images are from the people class although all contain human figures. The categories each image belongs to is indicated below the image.

remove this ambiguity. The point to be noted here is that although the human figures occupy a significant part of all of these images, none of them belong to the people class. In general, similar situations can be found in most classes where images contain objects of another class. One possible course of action for future works could be a fuzzy class membership for each image where typically an image is assigned multiple class labels in order of probability. That way, a man holding a gun would be classified both as a man and a gun which would be a more logical way to classify the images in this dataset.

4.4 Summary

The contributions of this chapter are in the generation of novel descriptors for object and scene image classification based on color, shape, spatial and local information, and Gabor wavelet transformation. In particular, a new GPHOG descriptor is created to improve upon the popular PHOG descriptor by encoding local, shape and spatial information of an image. Then the classification performance of the GPHOG descriptor is comparatively assessed in grayscale and six different color spaces – RGB, HSV, YCbCr, oRGB, DCS

and YIQ – and the robust YIQ-GPHOG and YCbCr-GPHOG features are further proposed that are effectively suitable for object and scene image classification. Finally, a new FC-GPHOG descriptor is presented by integrating the Principal Component Analysis (PCA) features of the GPHOG descriptors in the six different color spaces to further combine color, shape, local and wavelet-based features. Experimental results using two grand challenge datasets show that the proposed new FC-GPHOG descriptor outperforms the PHOG and also achieves an image classification performance better than or comparable to other popular image descriptors, such as the Scale Invariant Feature Transform (SIFT) based Pyramid Histograms of visual Words (PHOW) descriptor, the Color SIFT four Concentric Circles (C4CC), Spatial Envelope, and Local Binary Patterns (LBP).

CHAPTER 5

NOVEL COLOR GABOR-LBP-PHOG (GLP) IMAGE DESCRIPTORS

Chapter 4 introduced the GPHOG and the FC-GPHOG descriptors that improve upon the popular Pyramid of Histograms of Oriented Gradients (PHOG) descriptor for object and scene image classification. This chapter presents a new set of color descriptors that further encodes texture information along with shape, color and wavelet information from an image. To this end, first, the Gabor-LBP (GLBP) descriptor is derived by accumulating the Local Binary Patterns (LBP) histograms of all the component images produced by applying Gabor filters. Then, by combining the GPHOG and the GLBP descriptors using an optimal feature representation method, a novel Gabor-LBP-PHOG (GLP) image descriptor is proposed which performs well on different image categories. Finally, a Fused Color GLP (FC-GLP) feature is proposed by integrating the PCA features of the six color GLP descriptors. The Principal Component Analysis (PCA) and the Enhanced Fisher Model (EFM) are applied for feature extraction and the nearest neighbor classification rule is used for classification. The effectiveness of the proposed GLP and FC-GLP feature vectors for image classification is evaluated using three grand challenge datasets, namely the Caltech 256 dataset, the MIT Scene dataset and the UIUC Sports Event dataset.

5.1 Novel Gabor-based Color Image Descriptors

In this section, first the formation of the Gabor-LBP (GLBP) descriptor is described and then the creation of the new GLP descriptor is explained.

5.1.1 The Gabor-LBP (GLBP) Descriptor

The Gabor-LBP (GLBP) descriptor integrates the Gabor magnitude responses of an image and texture information by deriving their Local Binary Patterns (LBP).

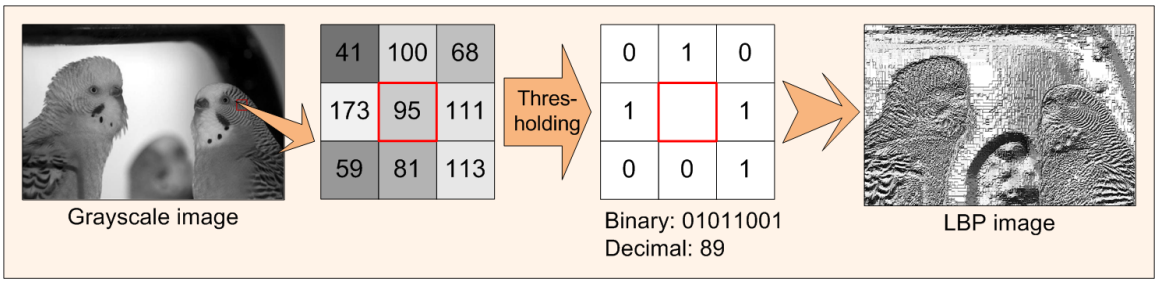


Figure 5.1 A grayscale image on the left, its Local Binary Patterns (LBP) image on the right, and the illustration of the computation of the LBP code for a center pixel with gray level 95.

The Local Binary Patterns (LBP) method derives the texture description of a grayscale image by comparing a center pixel with its neighbors (Ojala et al. 1994, 1996, 2002). In particular, for a 3×3 neighborhood of a pixel $\mathbf{p} = [x, y]^t$, $\mathbf{p}$ is the center pixel used as a threshold. The neighbors of the pixel $\mathbf{p}$ are defined as $N(\mathbf{p}, i) = [x_i, y_i]^t$, $i = 0, 1, \cdots, 7$, where i is the number used to label the neighbor. The value of the LBP code of the center pixel $\mathbf{p}$ is calculated as follows:

$$LBP(\mathbf{p}) = \sum_{i=0}^{7} 2^i S\{G[N(\mathbf{p}, i)] - G(\mathbf{p})\} \tag{5.1}$$

where $G(\mathbf{p})$ and $G[N(\mathbf{p}, i)]$ are the gray level of the pixel $\mathbf{p}$ and its neighbor $N(\mathbf{p}, i)$, respectively. S is a threshold function that is defined below:

$$S(x_i - x_c) = \begin{cases} 1, & \text{if } x_i \geq x_c \\ 0, & \text{otherwise} \end{cases} \tag{5.2}$$

LBP tends to achieve grayscale invariance because only the signs of the differences between the center pixel and its neighbors are used to define the value of the LBP code as shown in Eq. 5.1. Figure 5.1 shows a grayscale image on the left and its LBP image on the right. The two 3×3 matrices in the middle illustrate how the LBP code is computed for the center pixel whose gray level is 95. In particular, the center pixel functions as a threshold,

and after thresholding the right 3×3 matrix reveals the signs of the differences between the center pixel and its neighbors. Note that the signs are derived from Eqs. 5.1 and 5.2, and the threshold value is 95, as the center pixel is used as the threshold in the LBP definition. The binary LBP code is 01011001, which corresponds to 89 in decimal.

The Gabor-LBP descriptor is an extension of the LBP method in a way such that it incorporates not only texture information, but also wavelets and local cues from an image. In particular, the images are subjected to a series of Gabor filters defined in Eq. (3.4) in two scales and six orientations to get the magnitude responses. Then the LBP features are computed from each color component of the Gabor filtered images and the LBP histograms obtained from the Gabor filtered images are concatenated. The concatenated LBP histogram features are then finally normalized to zero mean and unit standard deviation to generate the GLBP descriptor. Figure 5.2 reveals the generation of the GLBP descriptor. More specifically, the first column shows a color image. The second column shows the Gabor filtered color images as a result of applying the series of Gabor filters to the color image, the third column displays the LBP histograms of the corresponding Gabor filtered images, and finally the fourth column shows the GLBP descriptor derived by normalizing the concatenated color LBP histograms of the Gabor filtered color images.

5.1.2 The GLP and the FC-GLP Descriptors

The GPHOG descriptor, presented in the preceding chapter, and GLBP descriptor encode local shape and texture information from Gabor wavelet responses of an image, respectively. To integrate the local, shape and texture cues extracted by these descriptors, the next descriptor is designed where the most expressive features of both the GLBP and the GPHOG feature vectors are integrated in cascade. The most expressive features are obtained by applying Principal component analysis (PCA). More specifically, the PCA features from the GLBP and the GPHOG vectors are taken and concatenated which are then normalized to zero mean and unit standard deviation, to form the Gabor-LBP-PHOG (GLP)

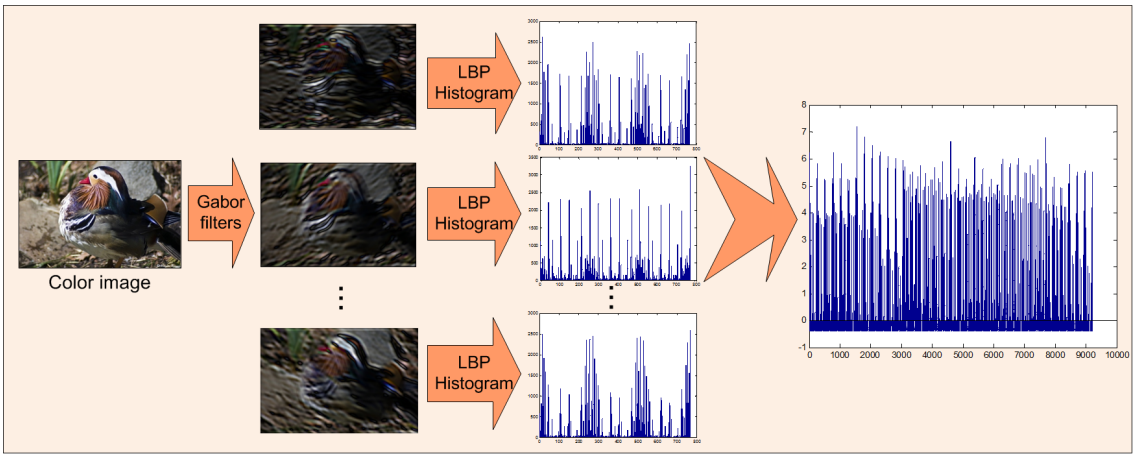


Figure 5.2 A color image, its Gabor filtered color images, the LBP histograms of the Gabor filtered color images, and the GLBP descriptor derived from the concatenation and subsequent normalization of the color LBP histograms of the Gabor filtered color images.

descriptor. It should be noted that for generating the GLBP and GPHOG vectors, the Gabor parameter values used are $\phi = 0$, $\sigma = 8$, $\gamma = 1$, ν = 1/16 and $\theta = [0, \pi/6, \pi/3, \pi/2, 2\pi/3, 5\pi/6]$.

The GLP feature vector is extended to six different color spaces, namely RGB, HSV, oRGB, YCbCr, YIQ and DCS as well as in grayscale. PCA is used for the optimal representation of the color GLP vectors with respect to minimum mean square error, and the PCA features of the six normalized color GLP descriptors are further combined to form the novel Fused Color GLP (FC-GLP) descriptor which outperforms the classification results of the individual color GLP features.

5.1.3 Classifier Used

Learning and classification is performed using Enhanced Fisher Linear Discriminant Model (EFM) (Liu and Wechsler 2000) and the nearest neighbor classification rule. Figure 5.3 gives an overview of multiple feature fusion methodology, the EFM feature extraction method, and the classification stages.

5.2 Experiments

In this section, first a brief description of the datasets that are used for the experiments is provided, and then the classification performance of the novel color GLP and FC-GLP descriptors is shown.

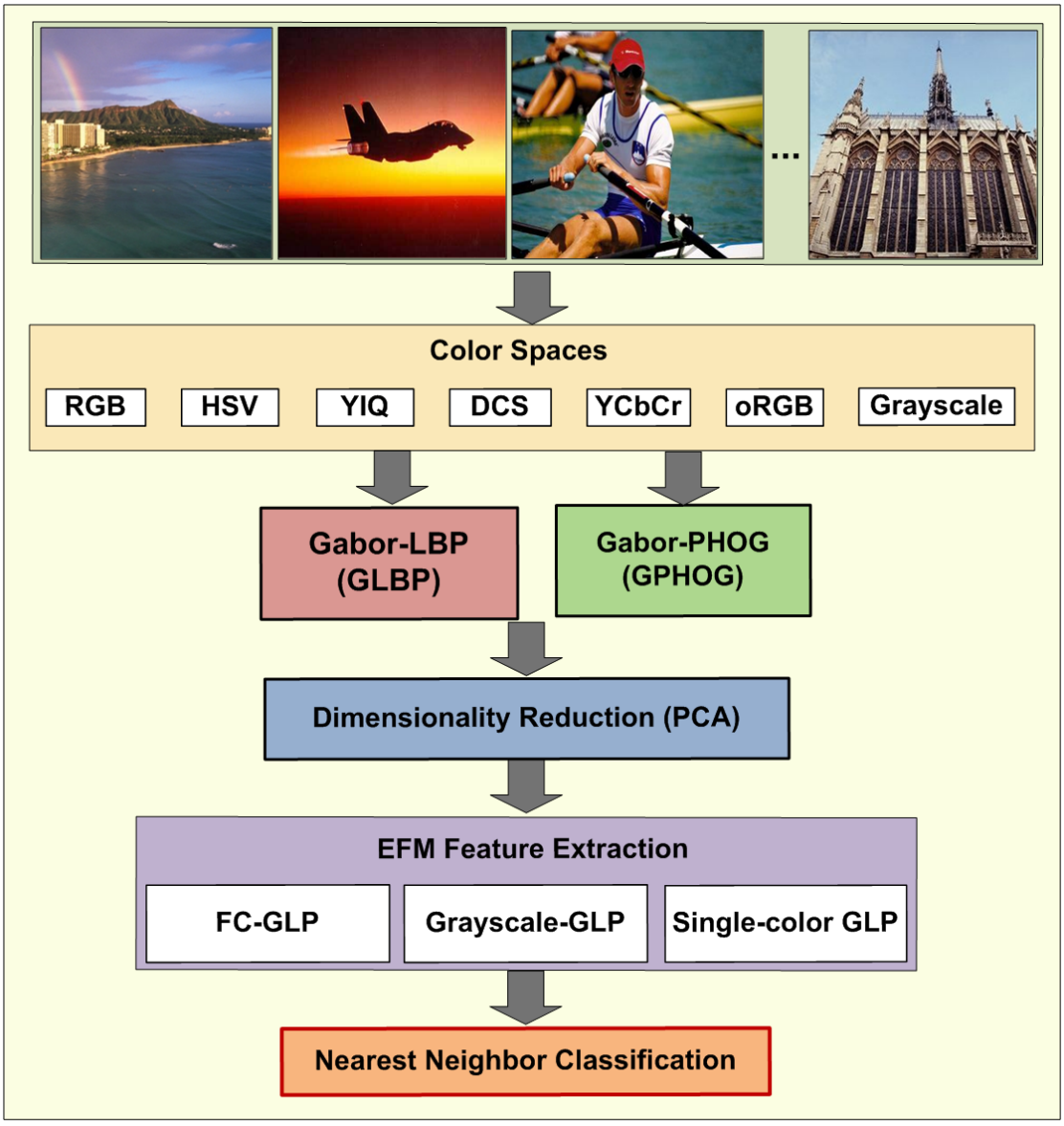


Figure 5.3 An overview of multiple features fusion methodology, the EFM feature extraction method, and the classification stages.

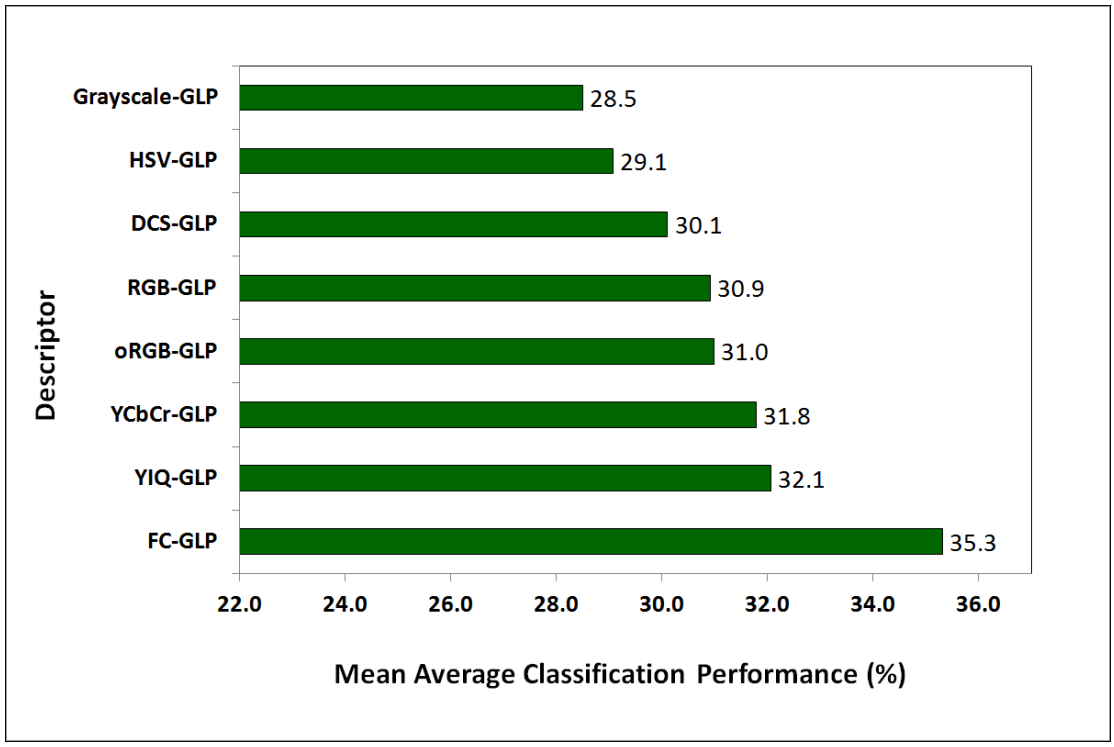


Figure 5.4 The mean average classification performance of the proposed color GLP and FC-GLP descriptors on the Caltech 256 dataset.

5.2.1 Datasets

The descriptors are tested using three popular and publicly available datasets, namely: the Caltech 256 dataset, the UIUC Sports Event dataset, and the MIT Scene dataset.

The Caltech 256 Dataset: The Caltech 256 dataset (Griffin et al. 2007) holds 30,607 images divided into 256 object categories and a clutter class. Section 3.3.1 contains detailed description of this dataset. Figure 3.3 displays some sample images from this dataset.

On this dataset, experiments are conducted using a protocol defined in (Griffin et al. 2007). For each class, 50 images are used for training and 25 images for testing, and five runs of experiments are done using the data splits that are provided on the Caltech website (Griffin et al. 2007).

The UIUC Sports Event Dataset: The UIUC Sports Event dataset (Li and Fei-

Table 5.1 Comparison of the Classification Performance (%) with Other Methods on Caltech 256 Dataset

Descriptor		Performance (%)
#train = 12800, #test = 6400		
oRGB-SIFT	(Verma et al. 2010)	23.9
gray-PHOW		25.9
color-PHOW		29.9
CSF	(Verma et al. 2010)	30.1
FC-GLP	**(Proposed)**	**35.3**
CGSF	(Verma et al. 2010)	35.6

Fei 2007) contains eight sports event categories: rowing (250 images), badminton (200 images), polo (182 images), bocce (137 images), snowboarding (190 images), croquet (236 images), sailing (190 images), and rock climbing (194 images). A few sample images of this dataset can be seen in Figure 3.5.

From each class, 70 images are used for training and 60 images for testing the classification performance of the descriptors, and this is done for five random splits. Other researchers (Bo et al. 2011; Li et al. 2010) have also reported using the same number of images for training and testing.

The MIT Scene Dataset: The MIT Scene dataset (Oliva and Torralba 2001) has 2,688 images classified as eight categories: 360 coast, 328 forest, 260 highway, 308 inside of cities, 374 mountain, 410 open country, 292 streets, and 356 tall buildings. A detailed description of this dataset is provided in Section 3.3.1. Figure 3.4 displays some sample images from this dataset.

From each class, 250 images are used for training and the rest of the images for testing the performance. A second set of experiments is also performed for this dataset using 100 training images from each class and the rest of the images for testing. For each of the experiments, a five-fold cross validation is done.

Table 5.2 Comparison of the Classification Performance (%) with Other Methods on the UIUC Sports Event Dataset

Descriptor		Performance (%)
#train = 560, #test = 480		
SIFT+GGM	(Li and Fei-Fei 2007)	73.4
OB	(Li et al. 2010)	76.3
gray-PHOW		76.4
CA-TM	(Niu et al. 2012)	78.0
color-PHOW		79.0
SIFT+SC	(Bo et al. 2011)	82.7
FC-GLP	**(Proposed)**	**84.3**
HMP	(Bo et al. 2011)	85.7

5.2.2 Results and Discussion

In this section, the performance of the proposed GLP and FC-GLP descriptors is evaluated in the three datasets, and also a comparison is made with some popular descriptors. Specif-

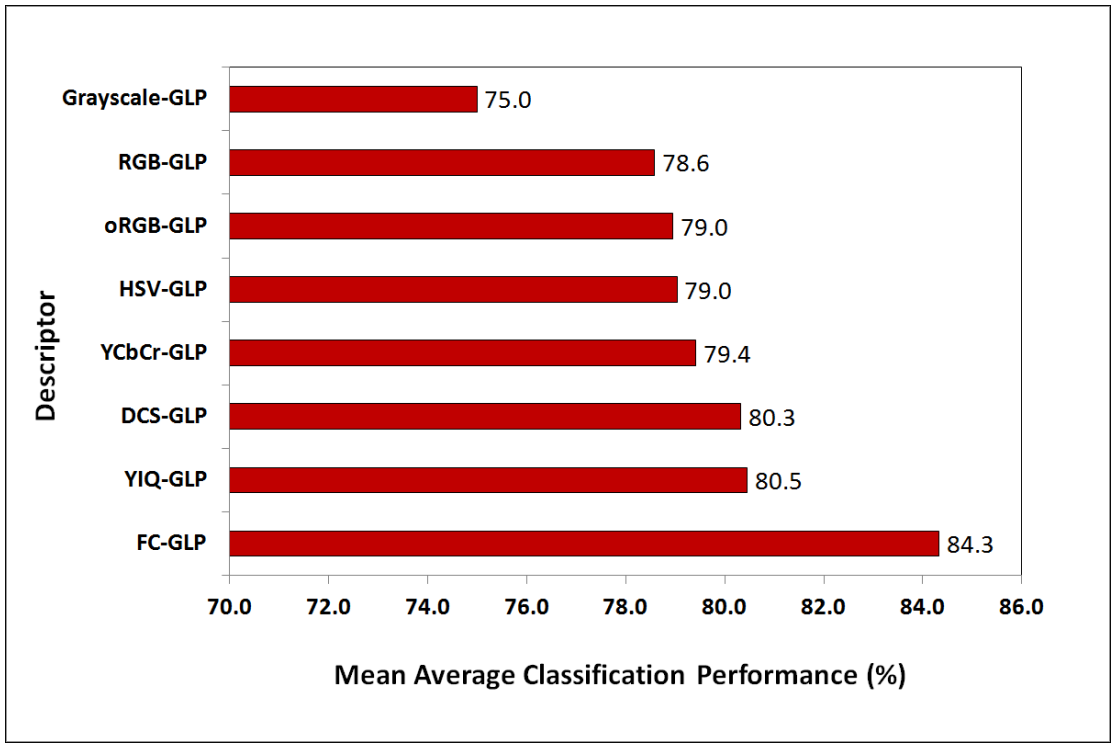


Figure 5.5 The mean average classification performance of the proposed GLP descriptor in individual color spaces as well as after fusing them on the UIUC Sports Event dataset.

ically, the FC-GLP descriptor is again compared with the popular SIFT-based Pyramid Histograms of visual Words (PHOW) descriptor (Bosch et al. 2007a) on all three datasets. To compare the proposed FC-GLP descriptor with the popular SIFT-based feature, the Pyramid Histograms of visual Words (PHOW) feature vector (Bosch et al. 2007a) is generated using the software package VLFeat (Vedaldi and Fulkerson 2010). For both PHOW and FC-GLP, PCA obtains the most expressive features and the EFM-NN classifier is employed in order to make a fair comparison. In addition, the classification performance achieved by the FC-GLP descriptor coupled with the EFM-NN classifier is also compared to the image classification performance of some other popular methods as reported in literature.

In the Caltech 256 dataset, YIQ-GLP performs the best among single-color descriptors giving 32.1% success followed by YCbCr-GLP and oRGB-GLP with 31.8% and 31.0% classification rates, respectively. Figure 5.4 shows the success rates of the GLP de-

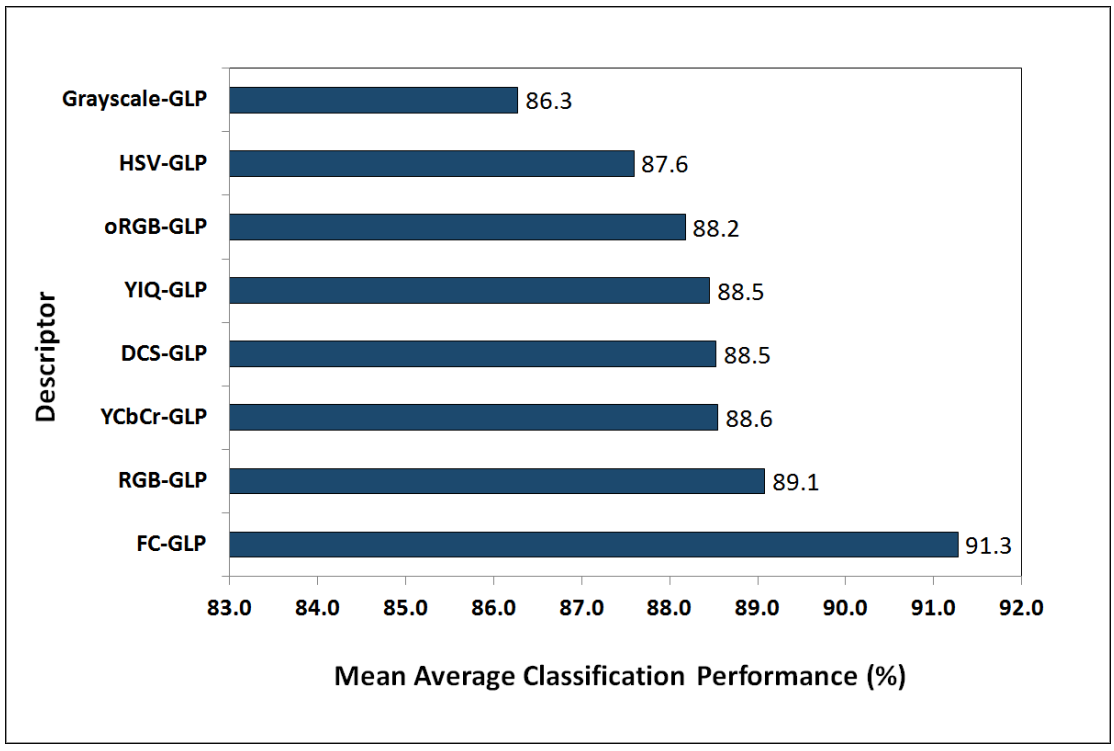


Figure 5.6 The mean average classification performance of the GLP descriptor in individual color spaces as well as after fusing them on the MIT Scene dataset.

Table 5.3 Comparison of the Classification Performance (%) with Other Methods on the MIT Scene Dataset

Descriptor		Performance (%)
#train = 800, #test = 1888		
CLF	(Banerji et al. 2011)	79.3
CGLF	(Banerji et al. 2011)	80.0
gray-PHOW		82.5
SE	(Oliva and Torralba 2001)	83.7
color-PHOW		84.3
CGLF+PHOG	(Banerji et al. 2011)	84.3
C4CC	(Bosch et al. 2006)	86.7
FC-GLP	**(Proposed)**	**87.5**
#train = 2000, #test = 688		
gray-PHOW		86.2
CLF	(Banerji et al. 2011)	86.4
CGLF	(Banerji et al. 2011)	86.6
color-PHOW		89.3
CGLF+PHOG	(Banerji et al. 2011)	89.5
FC-GLP	**(Proposed)**	**91.3**

scriptors for this dataset. The FC-GLP descriptor here achieves a success rate of 35.3%. Table 5.1 compares the results with other methods.

In the UIUC Sports Event dataset, the YIQ-GLP is the best single-color descriptor at 80.5% followed by DCS-GLP and YCbCr-GLP, respectively. The combined descriptor

Table 5.4 Category-wise GLP Descriptor Performance (%) on the UIUC Sports Event Dataset. Note that the Categories are Sorted on the FC-GLP Results

Category	FC	YIQ	DCS	YCbCr	HSV	oRGB	RGB	Grayscale
rock climbing	**96**	94	94	93	93	94	93	90
sailing	**94**	**94**	**94**	**94**	93	92	**94**	92
badminton	**93**	88	86	88	85	87	87	87
rowing	**88**	87	86	85	84	86	85	82
snow boarding	**88**	84	83	83	83	82	81	75
polo	**86**	76	76	78	81	74	78	72
croquet	**75**	74	71	69	65	67	68	60
bocce	**55**	47	53	46	48	50	43	42
Mean	**84.3**	**80.5**	**80.3**	**79.4**	**79.0**	**79.0**	**78.6**	**75.0**

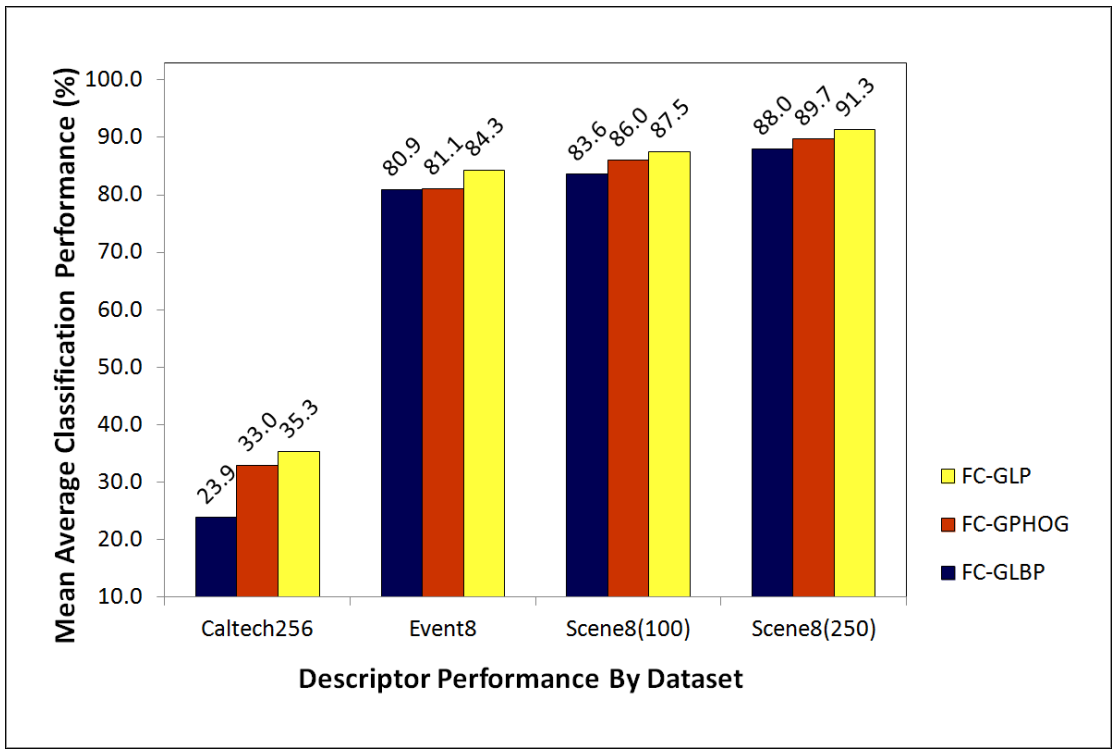


Figure 5.7 The comparative mean average classification performance of the FC-GLBP, FC-GPHOG and FC-GLP descriptors on the Caltech 256, UIUC Sports Event and MIT Scene (with 100 and 250 training images per class) datasets.

FC-GLP gives a mean average performance of 84.3%. See Figure 5.5 for details. Table 5.2 compares the result achieved by the proposed descriptor with that obtained by other

Table 5.5 Category-wise GLP Descriptor Performance (%) on the MIT Scene Dataset. Note that the Categories are Sorted on the FC-GLP Results

Category	FC	RGB	YCbCr	DCS	YIQ	oRGB	HSV	Grayscale
forest	**97**	**97**	96	96	96	96	**97**	96
highway	**94**	90	88	90	88	90	90	88
tall building	94	**95**	93	94	94	92	93	94
street	93	92	90	**94**	92	90	90	89
coast	**93**	89	**93**	88	91	91	90	87
mountain	**91**	87	87	88	89	86	87	72
inside city	88	**90**	86	87	84	87	83	86
open country	**80**	77	76	72	73	74	69	69
Mean	**91.3**	**89.1**	**88.6**	**88.5**	**88.5**	**88.2**	**87.6**	**86.3**

researchers. The category wise recognition performance of the GLP descriptors on this dataset is shown in table 5.4.

For the MIT Scene dataset, using 250 training images per class, the RGB-GLP is the best single-color descriptor at 89.1% followed closely by YCbCr-GLP and DCS-GLP. The combined descriptor FC-GLP gives a mean average performance of 91.3%. See Figure 5.6 for details. Table 5.3 shows a comparison with that of other methods. Table 5.5 shows the class wise classification rates for this dataset on applying the proposed GLP descriptors.

Figure 5.7 gives a comparison of the FC-GLBP, FC-GPHOG descriptors and their fusion (FC-GLP) for image classification in the three datasets used for the experiments. It should be noted that the generation time of the GPHOG and the GLBP features varies linearly with the number of pixels in the input image. It can be observed that the six color GLP features beat the recognition performance of the Grayscale-GLP descriptor which show information contained in color images can be significantly more useful than that in grayscale images for classification. Furthermore, the fusion of multiple color GLP descriptors (FC-GLP) achieves significant increase in the classification performance over individual color GLP descriptors, which implies that various color GLP descriptors are not completely redundant for image classification tasks.

5.3 Summary

Two new Gabor-based local, texture, shape and color feature extraction methods, namely the GLP and the FC-GLP are proposed that combines the GPHOG and the GLBP features using an optimal feature representation method such as PCA. The proposed descriptors exceed or achieve comparable performance to some of the best classification performances reported elsewhere. Experimental results carried out using three grand challenge datasets show that the FC-GLP descriptor improves classification performance over the GLBP and GPHOG descriptors and can be successfully applied for object and scene image classification.

CHAPTER 6

THE INNOVATIVE GABOR-LBP-HOG (GLH) DESCRIPTOR

In Chapter 5, the GLP descriptor, formed by fusing the GPHOG and the GLBP descriptors outperform both of them and is found to be promising for image classification tasks. Based on the same idea, this chapter proposes a fusion descriptor by integrating the GHOG and the GLBP descriptors and investigates its performance for image classification.

A novel set of color image descriptors is proposed in this chapter based on texture, shape and Gabor wavelets for object and scene image classification by combining the GHOG and the GLBP descriptors as a feature set. Next, a comparative assessment of the classification performance of the GLH descriptor is made in six different color spaces as well as in grayscale. Finally, a new Fused Color GLH (FC-GLH) descriptor is proposed for object and scene image classification by concatenating the GLH descriptors in the six different color spaces to further incorporate color information. Feature extraction for the proposed descriptors employ Principal Component Analysis (PCA) and Enhanced Fisher Model (EFM), and the nearest neighborhood is exploited for final classification. Experimental results using three benchmark datasets, the Caltech 256 object categories dataset, the MIT Scene dataset, and the UIUC Sports Event dataset show that the proposed new image descriptors achieve better image classification performance than other popular image descriptors.

6.1 The GLH and the FC-GLH Descriptors

The GHOG and GLBP descriptors proposed earlier in this dissertation encode local shape and texture information from Gabor wavelet responses of an image respectively. To integrate the local, shape and texture cues extracted by these descriptors, the Gabor-LBP-HOG (GLH) descriptor is designed. To derive the GHOG and GLBP descriptors, the images

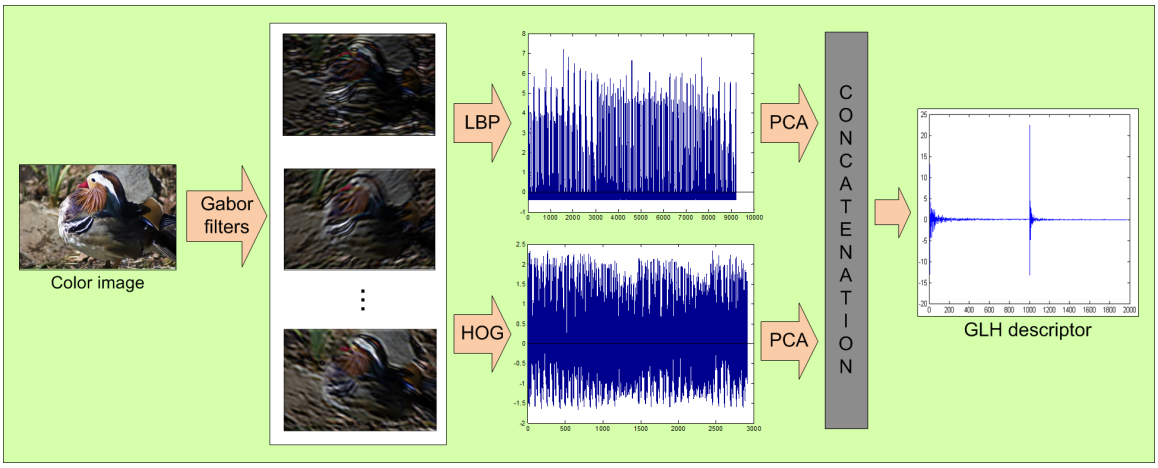


Figure 6.1 A color image, its Gabor filtered color images, the GLBP and the GHOG descriptors formed by applying LBP and HOG on the Gabor filtered color images respectively, the PCA and the concatenation process, and the GLH descriptor.

are first pre-processed by applying a series of Gabor filters in two scales and six different orientations. For all the experiments in this chapter, the chosen Gabor parameter values are $\sigma = 8$, $\gamma = 1$, $\nu = [1/8, 1/16]$, and $\theta = [0, \pi/6, \pi/3, \pi/2, 2\pi/3, 5\pi/6]$. The most expressive features of both the GLBP and the GHOG feature vectors are then integrated in cascade, where the most expressive features are obtained by applying Principal component analysis (PCA). The PCA technique has been reviewed in detail in Section 2.2. More specifically, the PCA features from the GLBP and the GHOG vectors are taken and concatenated which are then normalized to zero mean and unit standard deviation, to form the Gabor-LBP-HOG (GLH) descriptor. Figure 6.1 displays the generation of the GLH descriptor. It shows a color image in the first column, a series of Gabor filtered images in the second column produced as a result of applying twelve combinations of Gabor filters. The top and bottom rows of the third column show the GLBP and the GHOG descriptors respectively. The PCA and the concatenation process is revealed and then the GLH descriptor is shown in the last column.

To further incorporate the color information, the Fused Color GLH (FC-GLH) descriptor is presented by first computing the GLH in six color spaces and then concatenating

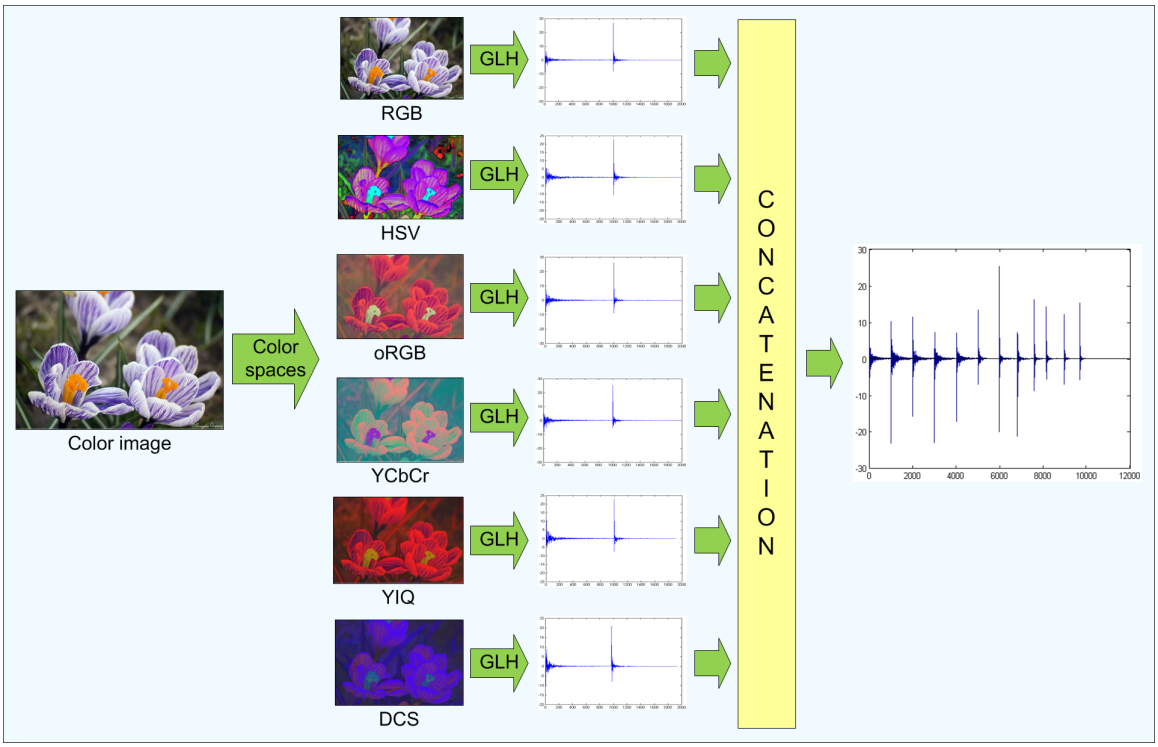


Figure 6.2 A color image, corresponding color images in the six color spaces, the GLH descriptors in the six color spaces, the concatenation process, and the FC-GLH descriptor.

the color GLH features. Figure 6.2 displays the creation of the FC-GLH descriptor. It shows a color image, its corresponding color images in the six color spaces, the six GLH descriptors of the images, the concatenation process, and the FC-GLH descriptor.

6.2 Classifier Used

The proposed new GLH and FC-GLH descriptors so formed are then tested for image classification performance by employing the EFM-NN classifier. This EFM-NN method has been reviewed in detail in Section 2.3. Figure 6.3 explains the process in detail.

6.3 Experiments

This section describes the datasets used and presents the experimental framework and classification results obtained by using the proposed descriptors.

6.3.1 Datasets

The descriptors are tested using three popular and publicly available datasets, namely: the Caltech 256 dataset, the UIUC Sports Event dataset, and the MIT Scene dataset.

The Caltech 256 Dataset: The Caltech 256 dataset (Griffin et al. 2007) holds 30,607 images divided into 256 object categories and a clutter class. Section 3.3.1 contains detailed description of this dataset. Figure 3.3 shows some sample images from this dataset.

On this dataset, experiments are conducted using a protocol defined in (Griffin et al. 2007). For each class, 50 images are used for training and 25 images for testing, and five runs of experiments are done using the data splits that are provided on the Caltech website (Griffin et al. 2007).

The UIUC Sports Event Dataset: The UIUC Sports Event dataset (Li and Fei-Fei 2007) contains eight sports event categories: rowing (250 images), badminton (200 images), polo (182 images), bocce (137 images), snowboarding (190 images), croquet (236 images), sailing (190 images), and rock climbing (194 images). Some sample images of this dataset can be seen in Figure 3.5.

Here, from each class, 70 images are used for training and 60 images for testing the classification performance of the GLH descriptors, and this is done for five random splits.

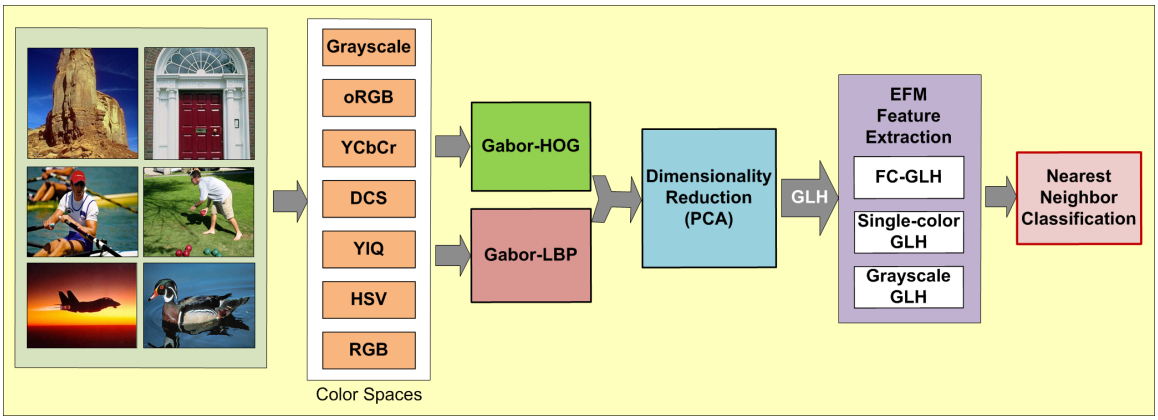


Figure 6.3 An overview of the formation of the grayscale GLH, the color GLH and the multiple features fusion (FC-GLH) methodology, the EFM feature extraction method, and the classification stages.

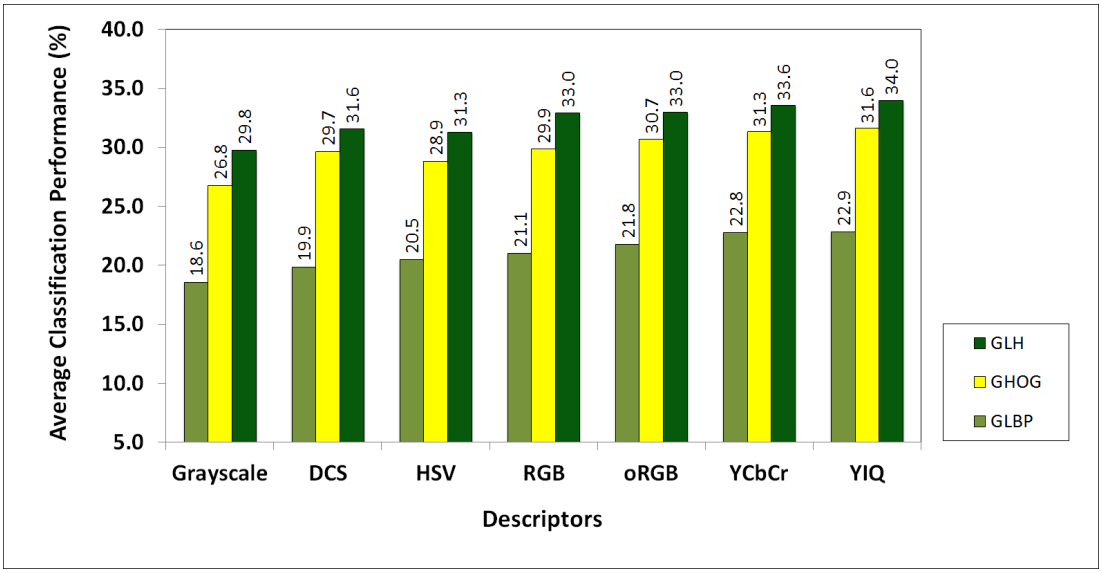


Figure 6.4 The average classification performance of the proposed GLBP, GHOG and GLH descriptors in the YIQ, the YCbCr, the oRGB, the RGB, the DCS, the HSV color spaces and also in grayscale using the EFM-NN classifier on the Caltech 256 dataset.

Other researchers (Bo et al. 2011; Li et al. 2010) have also reported using the same number of images for training and testing.

The MIT Scene Dataset: The MIT Scene dataset (Oliva and Torralba 2001) has 2,688 images classified as eight categories: 360 coast, 328 forest, 260 highway, 308 inside of cities, 374 mountain, 410 open country, 292 streets, and 356 tall buildings. A detailed description of this dataset is provided in Section 3.3.1. Figure 3.4 portrays some sample images from this dataset.

From each class, 100 images are used for training and the rest of the images for testing the performance. For each of the experiments, a five-fold cross validation is done.

6.3.2 Comparison of the GLH Descriptors in Different Color Spaces

In this section, a comparative assessment of the GLH descriptor is made in six different color spaces – RGB, HSV, oRGB, YCbCr, DCS and YIQ color spaces and in grayscale. The classification performance of the GLH descriptors are also compared with that of the

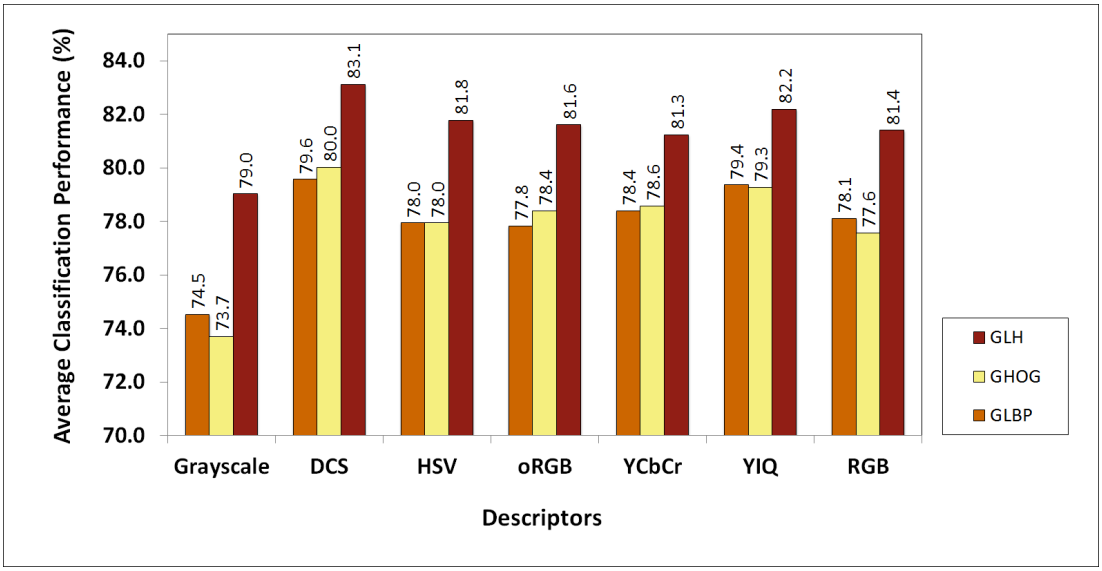


Figure 6.5 The average classification performance of the proposed GLBP, GHOG and GLH descriptors in the YIQ, YCbCr, oRGB, RGB, DCS, HSV color spaces and also in grayscale using the EFM-NN classifier on the UIUC Sports Event dataset.

GHOG and the GLBP descriptors.

Towards that end, the GLH descriptor is derived from each image in the different color spaces. Note that for some large-scale images, they are resized in such a way that the largest dimension does not exceed 400 pixels. Each input image is converted into grayscale as well as transformed into images in the six color spaces. Each image in a single color space first undergoes Gabor filtering in six orientations and two scales to produce twelve different Gabor-filtered images. The HOG and LBP descriptors are further computed from these Gabor filtered images and concatenated which are normalized to zero mean and unit standard deviation to finally derive the GHOG and GLBP features respectively. The PCA features of the GHOG and GLBP descriptors are fused to obtain the GLH descriptor.

Figure 6.4 shows the comparative classification performance of the proposed GLBP, GHOG and GLH descriptors in six different color spaces and also in grayscale on the Caltech 256 dataset. The horizontal axis shows the proposed descriptors in the six different color spaces and in grayscale, and the vertical axis denotes the average classification performance, which is the percentage of correctly classified images averaged across all the 256

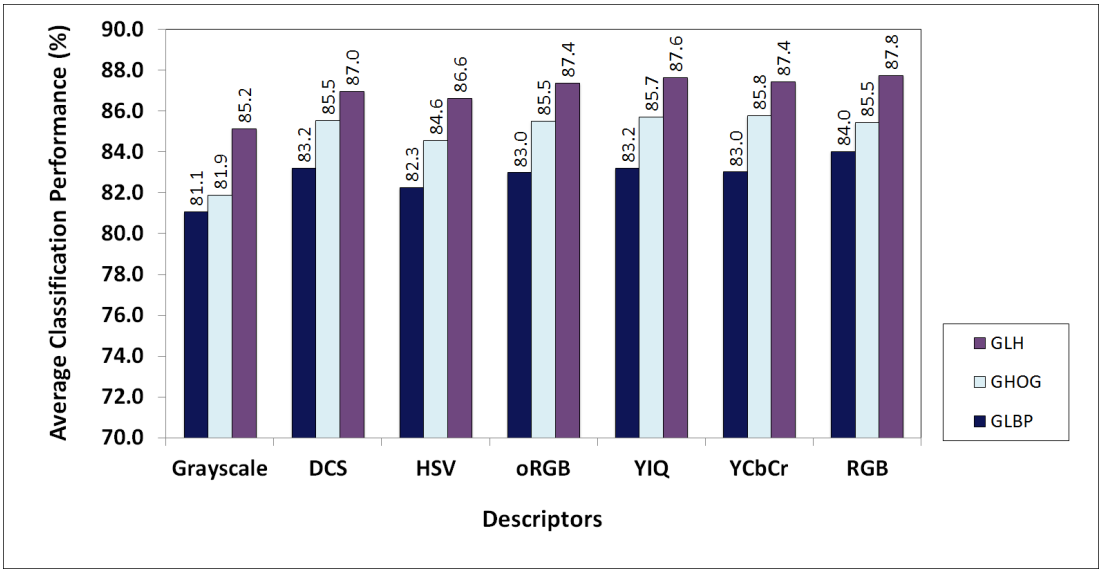


Figure 6.6 The average classification performance of the proposed GLBP, GHOG and GLH descriptors in the YIQ, YCbCr, oRGB, RGB, DCS, HSV color spaces and also in grayscale using the EFM-NN classifier on the MIT Scene dataset.

classes and the five runs of experiments. It shows that GLH descriptor in YIQ color space performs best with 34.0% classification performance. The classification performances of the GHOG and the GLBP descriptors in different color spaces are also shown in the figure for comparison.

For the UIUC Sports Event dataset, Figure 6.5 shows the detailed classification performance of the GLBP, GHOG and GLH descriptors in grayscale and in six different color spaces using the EFM-NN classifier. It can be seen from this figure that the GLH descriptor in DCS color space performs best with 83.1% classification rate.

On the MIT Scene dataset, the GLH descriptor performs fairly as well. Figure 6.6 shows the detailed classification performance of the GLBP, GHOG and GLH descriptors in grayscale and in six different color spaces using the EFM-NN classifier. Again, the horizontal axis shows the different descriptors in the different color spaces and in grayscale, and the vertical axis the average classification performance. Here, the GLH descriptor in RGB color space performs best with 87.8% classification rate.

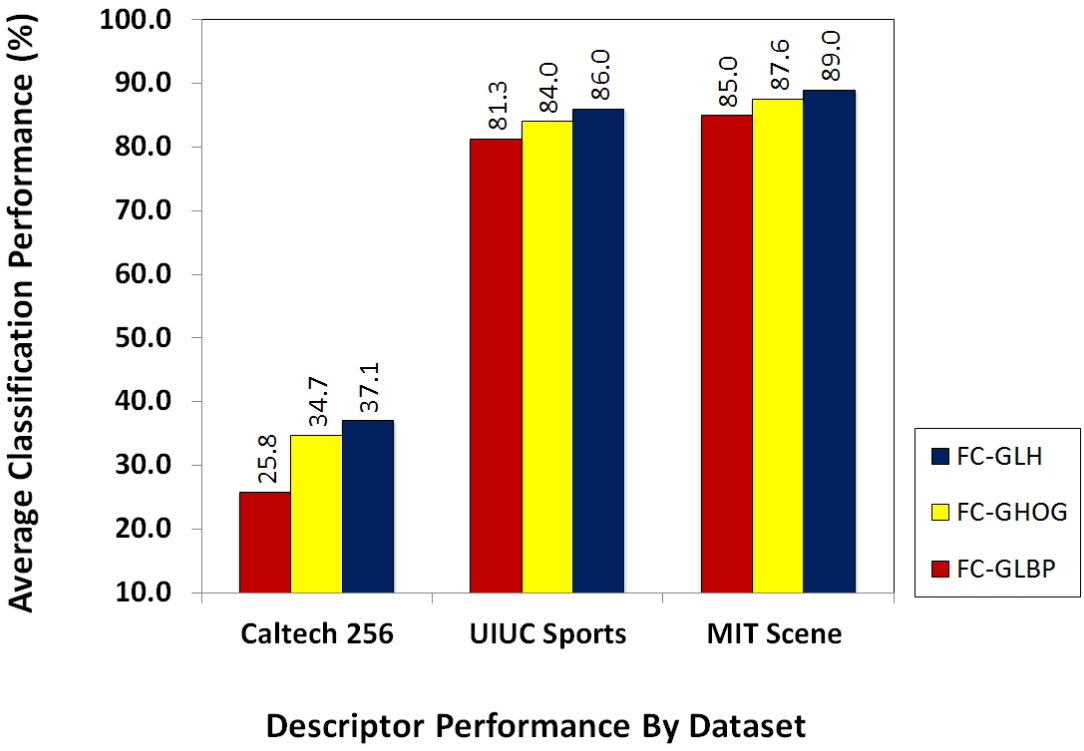


Figure 6.7 A comparison of the average classification performances of the FC-GLBP descriptor, the FC-GHOG descriptor, and the FC-GLH descriptor on the three image datasets. Note that all the three descriptors apply the EFM-NN classifier.

6.3.3 Comparison of the FC-GLH Descriptor and Some Other Methods

This section further evaluates the FC-GLH performance for image classification and compares it with that of some state-of-the-art descriptors.

Figure 6.7 reveals the comparison of the average classification performances of the FC-GLBP descriptor, the FC-GHOG descriptor, and the FC-GLH descriptor on the three image datasets. Note that the horizontal axis of this graph lists the three descriptors and the three datasets while the vertical axis shows the average classification performance as a percentage.

On the Caltech 256 dataset, the FC-GLH has an average classification performance of 37.1% which is better than both the FC-GHOG and the FC-GLBP descriptors with

34.7% and 25.8% classification success respectively. Table 6.1 shows the comparison of the classification performance of the proposed FC-GLH descriptor with that of other popular descriptors. In particular, on the Caltech 256 dataset, the FC-GLH descriptor achieves the average classification performance of 37.1%, compared to the color-PHOW and the gray-PHOW descriptors with the average classification rates of 29.9% and 25.9% respectively. It also outperforms the classification success achieved by oRGB-SIFT, Color Sift Fusion (CSF) and Color Grayscale Sift Fusion (CGSF) descriptors which yield 23.9%, 30.1% and 35.6% classification rates, respectively.

For the UIUC Sports Event dataset, the FC-GLH performs well achieving a classification performance of 86.0%. Table 6.2 reports the performance on this dataset. The FC-GLH performs better than FC-GHOG, FC-GLBP and other popular descriptors such as the PHOW, SIFT+GGM (Li and Fei-Fei 2007), Object Bank (OB) (Li et al. 2010), Context Aware Topic Model (CA-TM) (Niu et al. 2012), SIFC+SC and Hierarchical Matching Pursuit (HMP) (Bo et al. 2011) techniques.

On the MIT Scene dataset, the FC-GLH descriptor again gives the best classification performance of 89.0%, as compared to Color SIFT four Concentric Circles (C4CC) (Bosch et al. 2006) with a classification performance of 86.7%, to CGLF+PHOG (Banerji et al. 2011) with a classification performance of 84.3%, to Spatial Envelope (SE) with a classification performance of 83.7%, to Color LBP Fusion (CLF) (Banerji et al. 2011)

Table 6.1 Comparison of the Classification Performance (%) with Other Methods on Caltech 256 Dataset

Descriptor		**Performance (%)**
#train = 12800, #test = 6400		
oRGB-SIFT	(Verma et al. 2010)	23.9
gray-PHOW		25.9
color-PHOW		29.9
CSF	(Verma et al. 2010)	30.1
CGSF	(Verma et al. 2010)	35.6
FC-GLH	**(Proposed)**	**37.1**

Table 6.2 Comparison of the Classification Performance (%) with Other Methods on the UIUC Sports Event Dataset

Descriptor		Performance (%)
#train = 560, #test = 480		
SIFT+GGM	(Li and Fei-Fei 2007)	73.4
OB	(Li et al. 2010)	76.3
gray-PHOW		76.4
CA-TM	(Niu et al. 2012)	78.0
color-PHOW		79.0
SIFT+SC	(Bo et al. 2011)	82.7
HMP	(Bo et al. 2011)	85.7
FC-GLH	**(Proposed)**	**86.0**

Table 6.3 Comparison of the Classification Performance (%) with Other Methods on the MIT Scene Dataset

Descriptor		Performance (%)
#train = 800, #test = 1888		
CLF	(Banerji et al. 2011)	79.3
CGLF	(Banerji et al. 2011)	80.0
gray-PHOW		82.5
SE	(Oliva and Torralba 2001)	83.7
color-PHOW		84.3
CGLF+PHOG	(Banerji et al. 2011)	84.3
C4CC	(Bosch et al. 2006)	86.7
FC-GLH	**(Proposed)**	**89.0**

with a classification performance of 79.3%, and to Color Grayscale LBP Fusion (CGLF) (Banerji et al. 2011) with a classification performance of 80.0%. It also outperforms the PHOW, the FC-GLBP and the FC-GHOG descriptors. Table 6.3 shows the comparison on this dataset.

6.4 Summary

The contributions of this work are in the generation of novel descriptors for object and scene image classification based on color, texture, shape and Gabor wavelet transformation. In

particular, a novel Gabor-LBP-HOG (GLH) image descriptor is proposed which combines the GLBP and the GHOG descriptors to improve classification results. The GLH descriptor is generated in six different color spaces – RGB, HSV, YCbCr, oRGB, DCS and YIQ – as well as in grayscale. A new FC-GLH descriptor is also presented for object and scene image classification by integrating the GLH descriptors in the six different color spaces to further incorporate color information. Experimental results using three grand challenge datasets, the Caltech 256 object categories dataset, the MIT Scene dataset, and the UIUC Sports Event dataset show that the proposed new descriptors achieve better image classification performance than other popular image descriptors.

CHAPTER 7

NEW WIGNER-BASED LOCAL BINARY PATTERNS (WLBP) DESCRIPTOR

Chapters 3, 4, 5, 6 showed that using Gabor wavelets for designing the image descriptors achieved satisfactory classification results and enhanced performance. This chapter further explores wavelets to construct new descriptor. Also, the descriptors proposed in this dissertation in the previous chapters work on whole images. A recent literature survey shows that researchers have obtained promising results by using part-based approaches and visual bag of words methods. Since the advent of the bag of visual words model (Sivic and Zisserman 2003), there have been notable contributions to enhance recognition performance by developing new and robust image descriptors as well as effective classification frameworks that have resulted in reduced quantization loss and improved recall performance (Arandjelović and Zisserman 2013).

This chapter introduces a new local feature description method to categorize scene images. The problem of recognizing scene images is addressed by encoding local image information that can lead to an effective classification performance. To this end, the computationally efficient Local Binary Patterns (LBP) descriptor is first chosen that captures the variation in intensity between neighboring pixels to encode texture from images (Ojala et al. 1996, 1994). The LBP method has been found suitable for scene classification tasks (Banerji et al. 2011) and hence has been used alone or along with other features to develop new image descriptors (Sinha et al. 2012; Banerji et al. 2013). The Wigner distribution has been extensively used in signal processing. Based on the pseudo-Wigner distribution of images and the Local Binary Patterns (LBP) technique, four major contributions are made. First, a multi-neighborhood LBP for small image blocks is defined. Second, the multi-neighborhood LBP is combined with the pseudo-Wigner distribution of images for feature extraction. Third, the innovative WLBP feature vector is derived by utilizing the frequency domain smoothing, the bag-of-words model and spatial pyramid representations

of an image. Finally, extensive experiments are performed to evaluate the performance of the proposed WLBP descriptor. Specifically, the descriptor is tested for classification performance using a Support Vector Machine (SVM) classifier on three fairly challenging publicly available scene image datasets, namely the UIUC Sports Event dataset, the Fifteen Scene Categories dataset and the MIT Scene dataset. Experimental results reveal that the proposed WLBP descriptor outperforms the traditional LBP technique and yields results better than some other popular image descriptors.

7.1 Feature Description and Classification

This section first gives a brief review of the concepts used and then discusses the methodology adopted for developing the WLBP image descriptor.

7.1.1 Pseudo-Wigner Distribution

The Wigner distribution, also known as Wigner-Ville distribution is a generalized time-frequency representation proposed by Wigner (Wigner 1932) and Ville (Ville 1948) in 1932 and 1948 respectively. Although it has been extensively used in signal processing area, its applications in image processing are limited. Jacobson and Wechsler (Jacobson and Wechsler 1987) were the first researchers to apply the Wigner distribution to solve image processing problems. A family of Wigner distributions is called the pseudo-Wigner distribution (Vaidya and Haralick 1993).

In order to use the Wigner distribution function for image processing applications, it needs to be extended to two-dimensional space. Thus Wigner distribution of a two dimensional image is a four-dimensional distribution function which has two space domain variables and two frequency domain variables. The concept of windows is also applied here, which allows applying a sliding window to the original function in the time domain.

In this work, the pixel-wise pseudo-Wigner distribution for grayscale images has

been used, which is calculated with a N-pixels-one dimensional oriented square window where N is the operational window size (Gabarda and Cristóbal 2007). To compute the pixel-wise Wigner-distribution (W) of an image X, the algorithm takes an array of N pixels arranged in direction θ. For this work, the function has been chosen to be periodic which takes the $(N+1)$ pixel value to be equal to the value determined by the image in position $N=1$. Hence, for each pixel (i,j) of an image X, $W(i,j,k)$ is the pseudo-Wigner distribution of that pixel in the image, where $1 \leq k \leq N$. Only the first plane of W is chosen to design the proposed WLBP descriptor.

7.1.2 Local Binary Patterns (LBP)

The Local Binary Patterns (LBP) method encodes the texture features from a grayscale i.e. intensity image by comparing each pixel with its neighboring pixels (Ojala et al. 1994, 1996). Specifically, for a 3×3 neighborhood of a pixel $\mathbf{p} = [x,y]^t$, $\mathbf{p}$ is the center pixel used as a threshold. The neighbors of the pixel $\mathbf{p}$ are defined as $N(\mathbf{p},i) = [x_i,y_i]^t$, $i = 0,1,\cdots,7$, where i is the number used to label the neighbor. The value of the LBP code of the center pixel $\mathbf{p}$ is calculated as follows:

$$LBP(\mathbf{p}) = \sum_{i=0}^{7} 2^i S\{G[N(\mathbf{p},i)] - G(\mathbf{p})\} \tag{7.1}$$

where $G(\mathbf{p})$ and $G[N(\mathbf{p},i)]$ are the gray levels of the pixel $\mathbf{p}$ and its neighbor $N(\mathbf{p},i)$, respectively. S is a threshold function that is defined below:

$$S(x_i - x_c) = \begin{cases} 1, & \text{if } x_i \geq x_c \\ 0, & \text{otherwise} \end{cases} \tag{7.2}$$

LBP has been reviewed in detail in Section 5.1.1 and explained in Figure 5.1.

7.1.3 Sampling and Bag of Features

In order to derive the WLBP descriptor, first the image is sampled. Popular descriptors like SIFT (Lowe 2004) use multiscale keypoint detectors such as Laplacian of Gaussian or Harris-affine to select regions of interest within the image. This sampling method is appropriate for object recognition, but it has been found that dense sampling often outperforms the keypoint-based sampling methods (Nowak et al. 2006). This is particularly true of images with large uniform regions, where SIFT does not detect any keypoints. Scene images, such as the ones used for this work, often have such homogeneous regions depicting the sky or walls. For this purpose, a dense sampling approach is used in which the image is divided into a number of equal sized overlapping square blocks or patches using a uniform grid and each block is used as a separate region for extracting features. The scene images are sampled using 40×40 pixel overlapping blocks, each block offset by 10 pixels from the next. Such patches are extracted from all training images and then the patches are clustered to form visual words. This process is explained in Figure 7.1. The image shown on the left

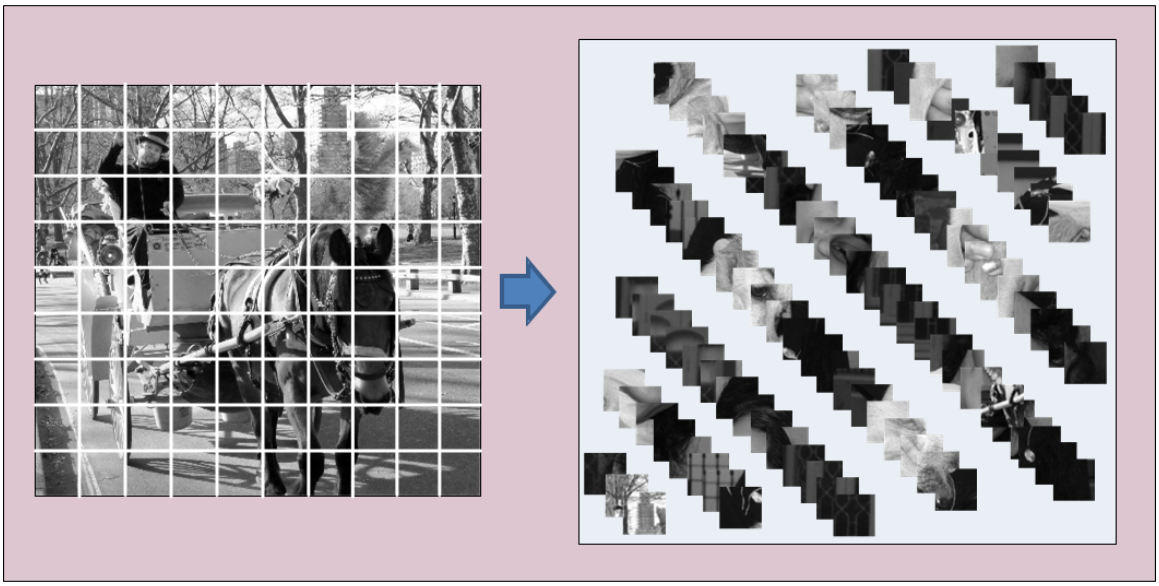


Figure 7.1 For the bag-of-words representation, a grayscale image is broken down into small image patches using a regular grid. This is called dense sampling. Overlapping patches are used for more accuracy.

is divided into uniform image patches by the regular grid displayed overlaid on the image, to form the image patches shown on the right. Figure 7.2 demonstrates the formation of visual words from an image after clustering the small image patches.

7.1.4 Multi-Scale WLBP Features for Small Image Patches

Now the feature extraction of the sampled image regions is discussed. First, the pixel-wise pseudo-Wigner distribution for each of the small image patches is computed as described in Section 7.1.1 in three different directions. For the experiments, the parameter values $N = 2$, $\theta = 0, \pi/4, \pi/2$ are used, and only the first planes of each of the three Wigner distributions have been retained for the image blocks for subsequent feature extraction.

The multi-neighborhood LBP features are then extracted from the image patch and the three images produced as a result of applying the Wigner-distribution on it. Different researchers have chosen various neighborhoods of different styles for extracting LBP

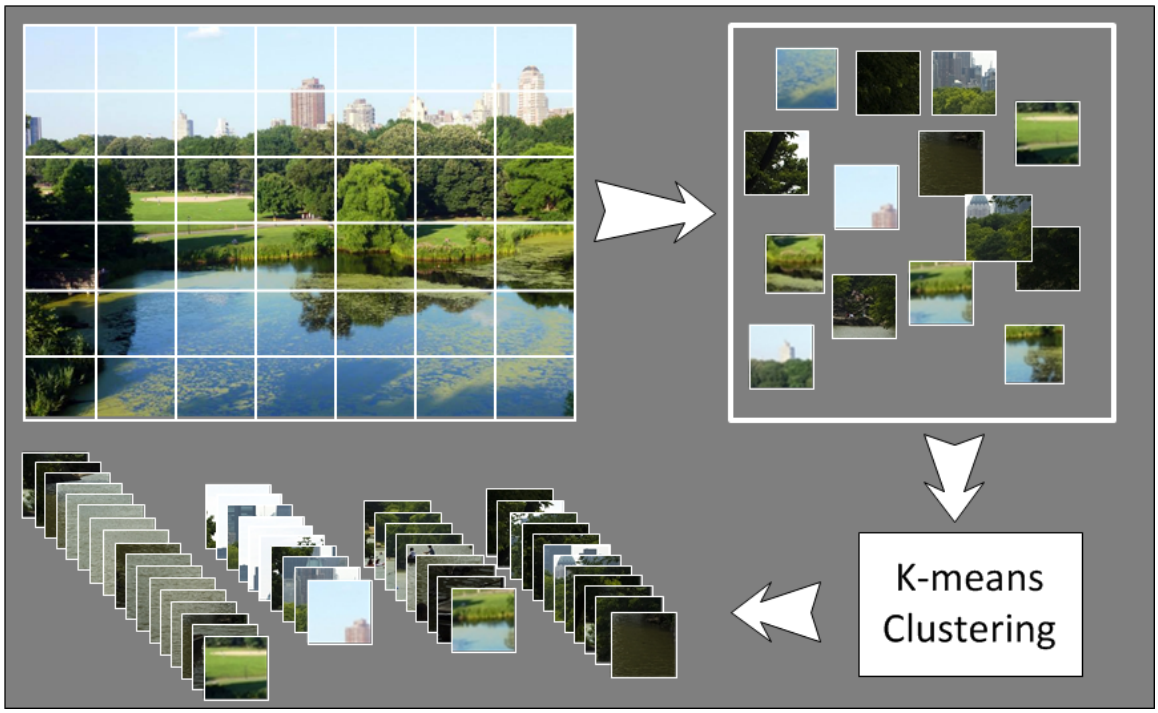


Figure 7.2 Formation of visual words from image patches using a popular clustering method.

features from an image (Zhu et al. 2010; Banerji et al. 2011; Gu and Liu 2013). The conventional 8-neighborhood LBP mask assigns one out of 2^8 possible intensity values to each pixel, resulting in a 256-bin histogram. However, since the image patches are small, 4-pixel neighborhood LBP masks are chosen to reduce the sparseness of the features. These LBP masks produce a dense 16-bin histogram, and eight such histograms from different neighborhoods and four sub-images are fused to design the 128-dimensional WLBP feature vector describing each image block. Figure 7.3 depicts the two 4-pixel neighborhoods used for generating the multi-neighborhood LBP descriptor used here.

The Discrete Cosine Transform (DCT) is a well-known technique of transforming an image to the frequency domain for various applications like compression, smoothing, etc. (Hafed and Levine 2001), where an image is decomposed into a combination of various uncorrelated frequency components. Specifically, the DCT of an image with the spatial resolution of $M \times N$, $f(x,y)$, where $x = 0, 1, \cdots, M-1$ and $y = 0, 1, \cdots, N-1$, transforms the image from the spatial domain to the frequency domain (Gonzalez and Woods 2008). DCT is thus able to extract the features in the frequency domain to encode different image

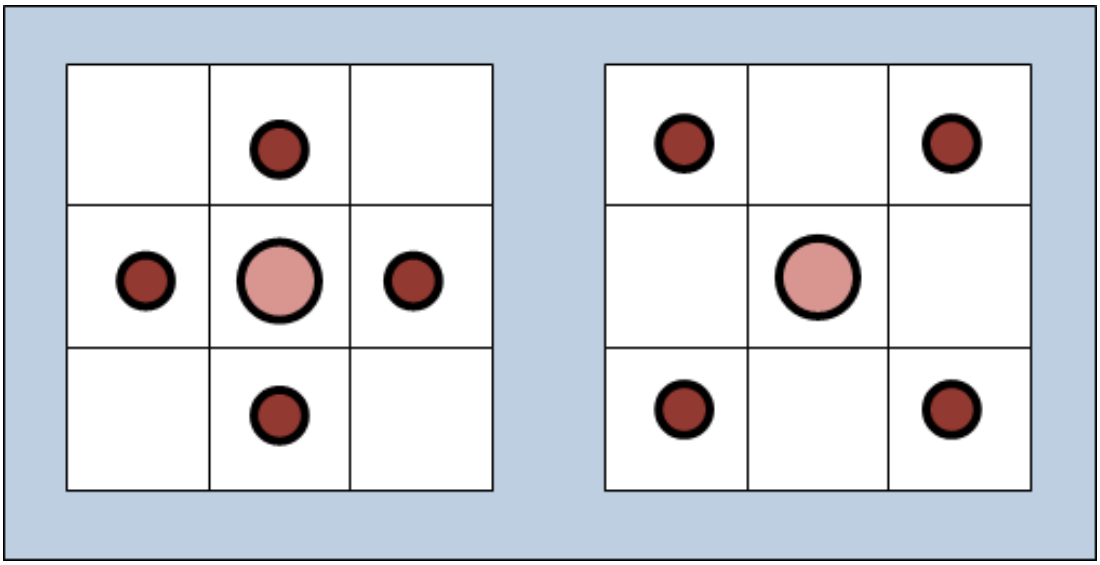


Figure 7.3 The two 4-neighborhood LBP masks used for computing the proposed WLBP descriptor.

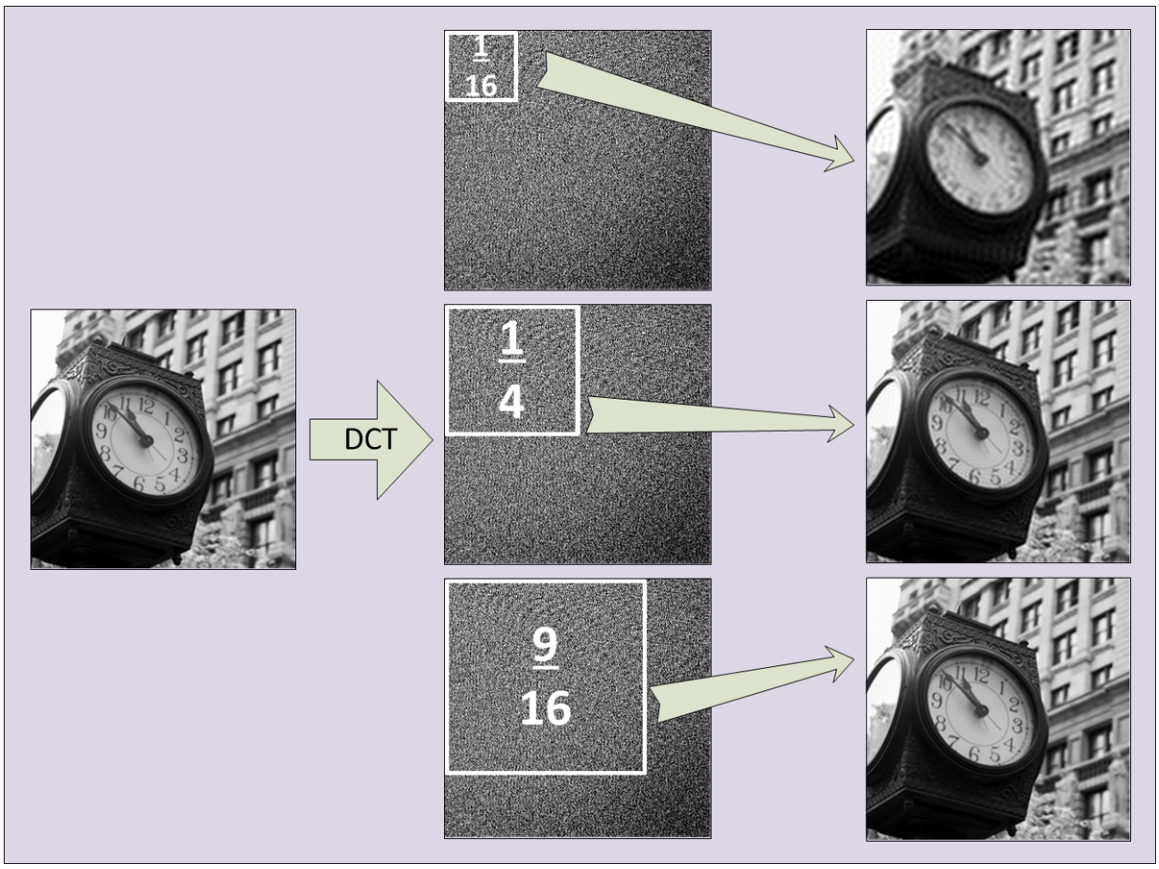


Figure 7.4 DCT can be used for smoothing out the image. The original image is transformed to the frequency domain and the lowest 1/16, 1/4 and 9/16 parts are used for regenerating the image, respectively, resulting in three output images with various degrees of smoothing.

details that are not directly accessible in the spatial domain. Due to these specific properties, DCT has been successfully applied to face recognition (Liu and Liu 2008; Chen et al. 2006; Hafed and Levine 2001). In the proposed method, DCT is used to eliminate higher frequencies from an image, resulting in a form of smoothing. To achieve image smoothing for capturing textures at different scales, the DCT technique is performed to transform the original image to frequency domain and the lowest 6.25%, 25% and 56.25% of frequencies are used to regenerate the image. This process is explained in Figure 7.4. The original image and the three images thus formed undergo the same process of dense sampling and WLBP feature extraction. All these features together form a bag of features, as shown in

Figure 7.5, that needs to be clustered into distinct visual words to form a visual vocabulary. Figure 7.6 illustrates the complete process of generating the WLBP features from a grayscale image.

7.1.5 Quantization and Pyramid Representation

The next stage is to quantize the bag of WLBP features extracted from the training images into a visual vocabulary with discrete visual words. For this step, the popular K-means algorithm is used. The vocabulary size used by researchers vary from a few hundreds (Lazebnik et al. 2006; Zhang et al. 2007) to several thousands (Sivic and Zisserman 2003; Zhao et al. 2006). In this work, the experiments have been performed with vocabularies of varying sizes and empirically a 1000-word vocabulary is chosen. After the creation of the visual vocabulary, each scene image is represented by a histogram of visual words. This is explained in Figure 7.7(a).

The image pyramid representation proposed by (Lazebnik et al. 2006) allows a

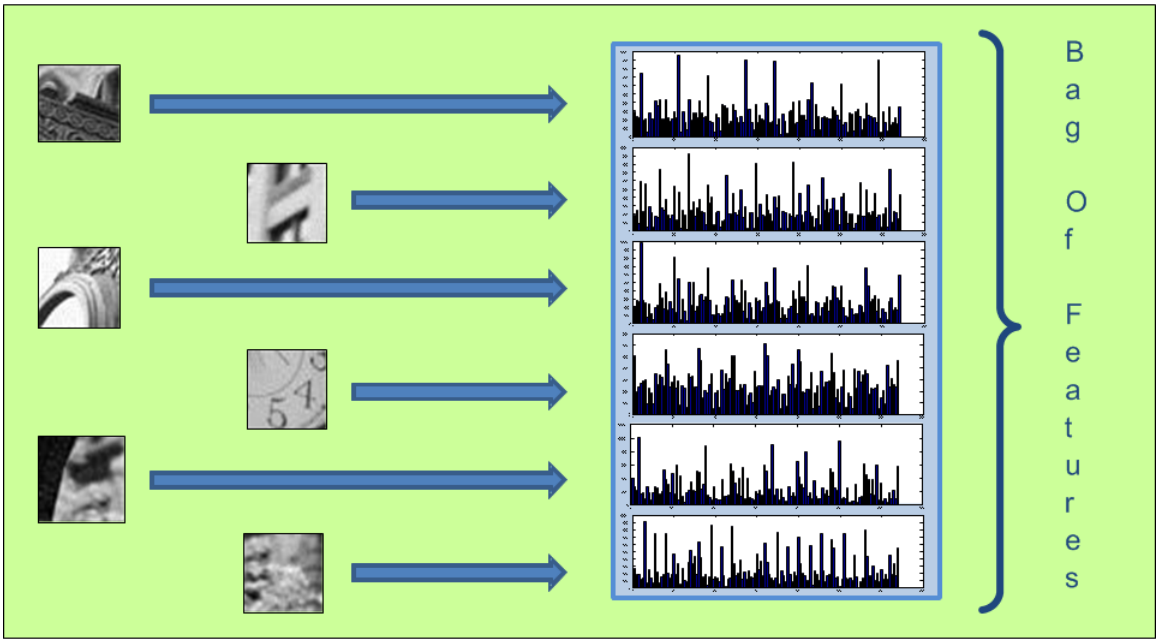


Figure 7.5 The features are computed from a large number of image patches from all training images and form a bag of features from which a visual vocabulary can be created.

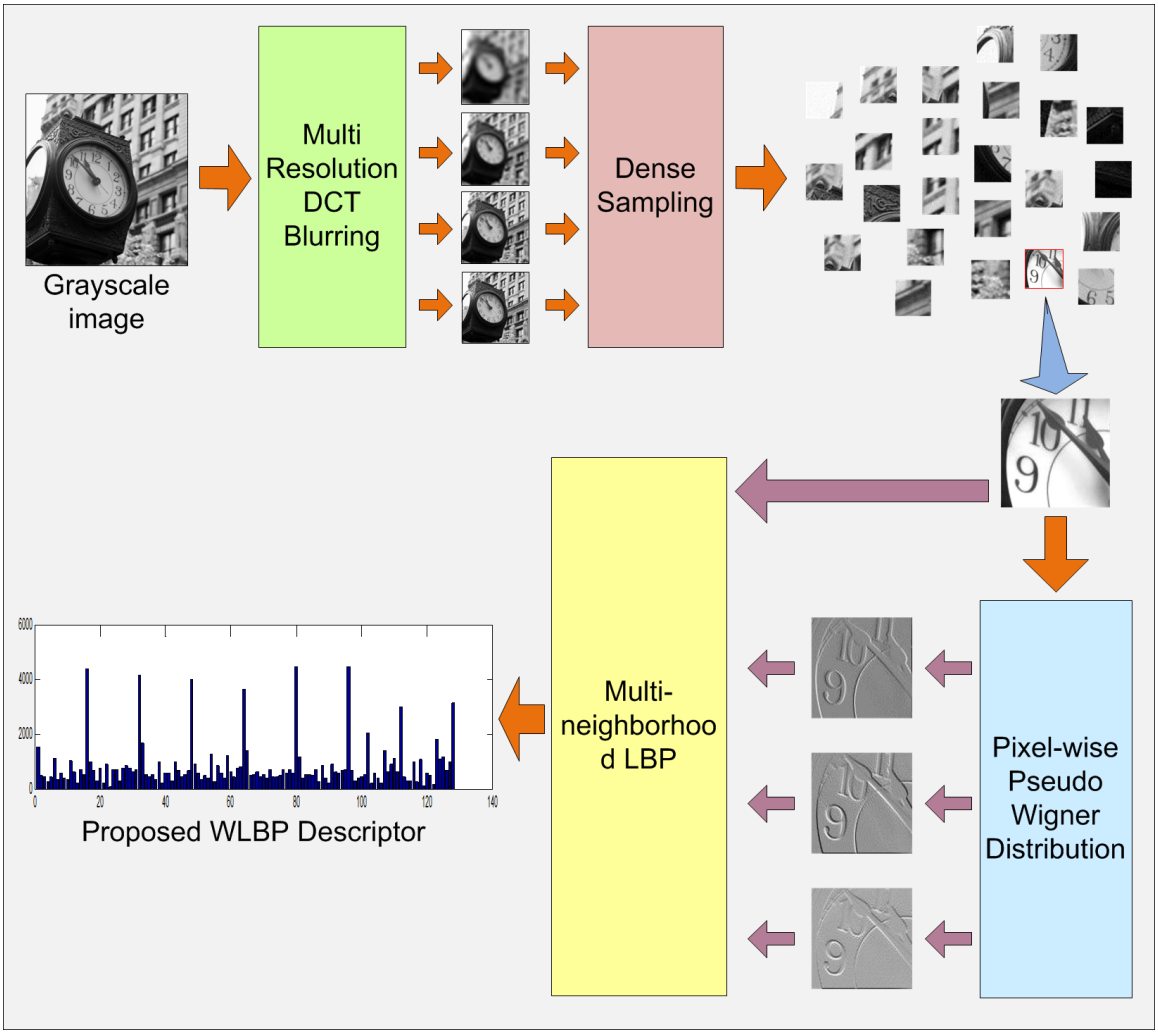


Figure 7.6 The process of computing the proposed WLBP descriptor has been simplified in this schematic diagram.

descriptor to represent local image features and their spatial layout. Here, at each level, an image is tiled into its successively smaller blocks and the feature vectors are computed for each block. These features from each pyramid level are then weighted accordingly, which are finally concatenated to form a pyramid histogram. This technique is explained in Figure 7.7(b). It should be noted that the histograms shown in Figure 7.7 are for illustration purposes only. For this work, only the second level of this pyramid has been utilized to keep the computational complexity low. Finally, a 4000 dimensional feature vector is constructed for each image.

7.1.6 Classifier Used

After all training and test images have been processed and the feature vectors have been generated, an SVM classifier is used for classification. SVM has been reviewed in detail in Section 2.4. It is a known fact in texture and other image classification that for comparing histograms, using χ^2 or Hellinger distance measures usually yields better results than Euclidean distance (Arandjelović and Zisserman 2012) . The use of the Hellinger kernel has been shown to benefit SIFT (Arandjelović and Zisserman 2012). Since the proposed WLBP descriptor is also a histogram, intuitively it seems that it should yield better classification results with the Hellinger kernel and it is empirically seen that using the Hellinger kernel does indeed improve the classification results greatly.

If x and y are n-vectors with unit Euclidean norm ($|x|_2 = 1$), then the Euclidean distance $d_E(x,y)$ between them is related to their similarity (kernel) $S_E(x,y)$ as

$$d_E(x,y)^2 = |xy|_2^2 = |x|_2^2 + |y|_2^2 - 2x^t y = 2 - 2S_E(x,y) \tag{7.3}$$

where $S_E(x,y) = x^t y$, and the last step follow from $|x|_2^2 = |y|_2^2 = 1$. The Euclidean similarity/kernel here needs to be replaced by the Hellinger kernel.

The Hellinger kernel, which is also known as the Bhattacharyya's coefficient, is defined for two L1 normalized histograms, x and y (i.e. $\sum_{i=1}^{n} x_i = 1$ and $x_i \geq 0$) as:

$$H(x,y) = \sum_{i=1}^{n} \sqrt{x_i y_i} \tag{7.4}$$

Arandjelović et al. suggest a simple algebraic manipulation to compare SIFT vectors by a Hellinger kernel (Arandjelović and Zisserman 2012). Since WLBP vectors are also based on histograms of words, the same technique can be applied to the WLBP vectors as well. This can be done in two steps: (i) L1 normalize the WLBP vector (originally it has unit L2 norm); (ii) square root each element. It then follows that $S_E(\sqrt{x}, \sqrt{y}) =$

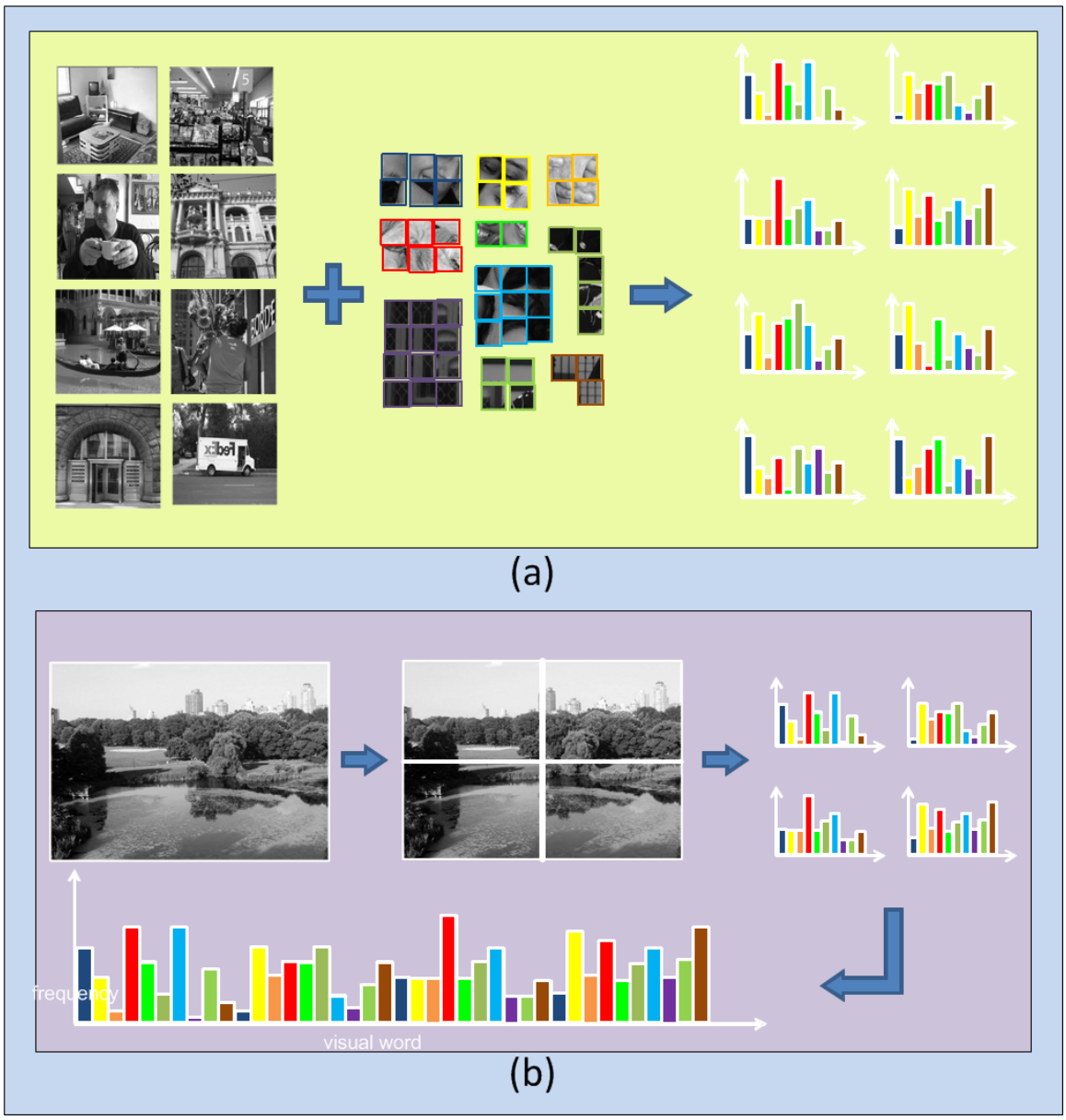


Figure 7.7 (a) All images are converted to histograms of visual words using the visual vocabulary created from the training images. (b) For the spatial pyramid representation, a full image is broken down into multiple spatial tiles. Then histograms of visual words are computed from each tile and concatenated.

$\sqrt{x}^t\sqrt{y} = H(x,y)$, and the resulting vectors are L2 normalized since $S_E(\sqrt{x},\sqrt{y}) = \sum_{i=1}^{n} = 1$ (Arandjelović and Zisserman 2012).

The key point is that comparing the square roots of the WLBP descriptors using Euclidean distance is equivalent to using the Hellinger kernel to compare the original WLBP

vectors:

$$d_E(\sqrt{x},\sqrt{y})^2 = 2 - 2H(x,y) \tag{7.5}$$

For the classification process, an SVM is trained independently for each class (one-vs-all classification). This is repeated for each category separately and the precision rates from all the iterations give the average precision which is the mean classification accuracy. A similar configuration has been successfully used by other researchers like (Sanchez et al. 2012) in recent works. The SVM implementation used here is the one that is distributed with the VLFeat package (Vedaldi and Fulkerson 2010).

7.2 Experiments

This section first introduces the three scene image datasets used for evaluating the classification performance of the WLBP descriptor, and then makes a comparative assessment of the classification performances of the LBP and the WLBP descriptors. Finally, the classification performance of the WLBP descriptor is compared with that of some popular image descriptors used by other researchers on these datasets. It should be noted that the results of other researchers are reported directly from their published work.

7.2.1 Datasets Used

Three publicly available and widely used image datasets are used in this work for assessing the classification performance of the proposed descriptor.

The UIUC Sports Event Dataset: The UIUC Sports Event dataset (Li and Fei-Fei 2007) contains 1,574 images from eight sports event categories: 250 rowing, 200 badminton, 182 polo, 137 bocce, 190 snowboarding, 236 croquet, 190 sailing, and 194 rock climbing. A detailed description of this dataset is provided in Section 3.3.1. Some sample images from this dataset are displayed in Figure 3.5.

The MIT Scene Dataset: The MIT Scene dataset (also known as OT Scenes)

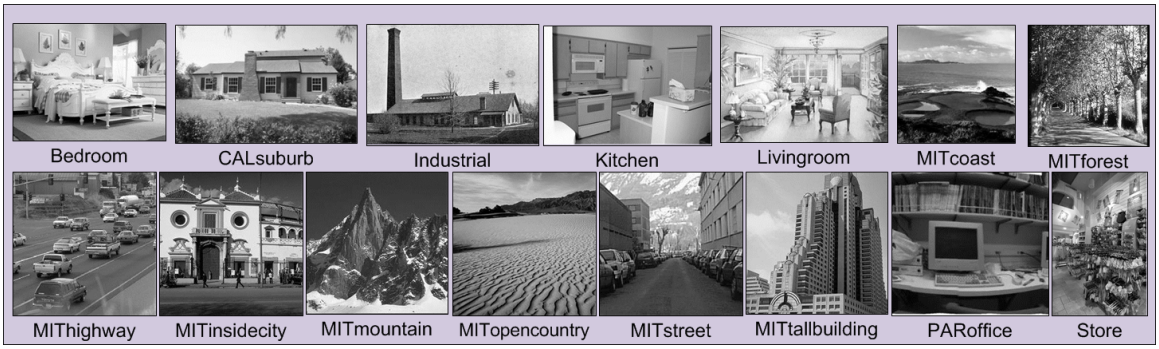


Figure 7.8 Some sample images from the Fifteen Scene Categories dataset.

(Oliva and Torralba 2001) has 2,688 images classified as eight categories: 360 coast, 328 forest, 260 highway, 308 inside of cities, 374 mountain, 410 open country, 292 streets, and 356 tall buildings. Section 3.3.1 provides a detailed description of this dataset. Figure 3.4 shows a few sample images from this dataset.

The Fifteen Scene Categories Dataset: The Fifteen Scene Categories dataset (Lazebnik et al. 2006) is composed of 15 scene categories: thirteen were provided by (Fei-Fei and Perona 2005), eight of which were originally collected by (Oliva and Torralba 2001) as the MIT Scene dataset, and two were collected by (Lazebnik et al. 2006). Each category has 200 to 400 images, most of which are grayscale. Figure 7.8 shows a few images from this dataset.

7.2.2 Comparison of the LBP, WLBP and Other Popular Descriptors

The classification performance of the proposed WLBP descriptor is now evaluated by comparing it with the traditional LBP feature and some other popular image descriptors on the three scene image datasets. To that end, first the WLBP feature vector is derived from each image in the dataset. To compute the WLBP descriptor, first each color image is converted to grayscale and then all the training images are divided into overlapping uniform image patches. Please note that the large scale images are resized in such a way that their largest dimension does not exceed 256 pixels. The WLBP features are extracted from all the im-

age patches generated from the grayscale image and the three DCT-smoothed images to generate a bag of features which is quantized using the K-means algorithm to form a visual vocabulary with 1000 words. Next, each training and test image is represented as a pyramid histogram of these visual words. An SVM classifier with a Hellinger kernel (Vapnik 1995; Vedaldi and Fulkerson 2010) is used for evaluating the relative classification performances of the LBP and the WLBP descriptors.

For the UIUC Sports Event dataset, 70 images from each class are used for training and 60 from each class for testing both the LBP and the WLBP descriptors. The results are obtained using five random splits of data where there is no overlap between the training and testing images of the same split. Figure 7.9 shows the relative average precisions achieved by the LBP and the WLBP descriptors on this dataset. Note that here, the horizontal axis shows the two descriptors and the three datasets, and the vertical axis shows the classification performance measured by average precision as percentage. Here, the WLBP descriptor outperforms the LBP by over 14%. The proposed WLBP vector also produces better results than other SIFT-based and state-of-the-art methods on this dataset, which is listed in Table 7.1.

Table 7.1 Comparison of the Classification Performance (%) of the Proposed Grayscale WLBP Descriptor with Other Popular Methods on the Three Image Datasets

Method		**UIUC Sports**	**MIT Scene**	**15 Scenes**
SIFT+GGM	(Li and Fei-Fei 2007)	73.4	-	-
OB	(Li et al. 2010)	76.3	-	-
KSPM	(Yang et al. 2009)	-	-	76.7
KC	(Van Gemert et al. 2010)	-	-	76.7
CA-TM	(Niu et al. 2012)	78.0	-	-
ScSPM	(Yang et al. 2009)	-	-	80.3
SIFT+SC	(Bo et al. 2011)	82.7	-	-
SE	(Oliva and Torralba 2001)	-	83.7	-
HMP	(Bo et al. 2011)	85.7	-	-
C4CC	(Bosch et al. 2006)	-	86.7	-
WLBP+SVM	**(Proposed)**	**86.2**	**92.2**	**85.1**

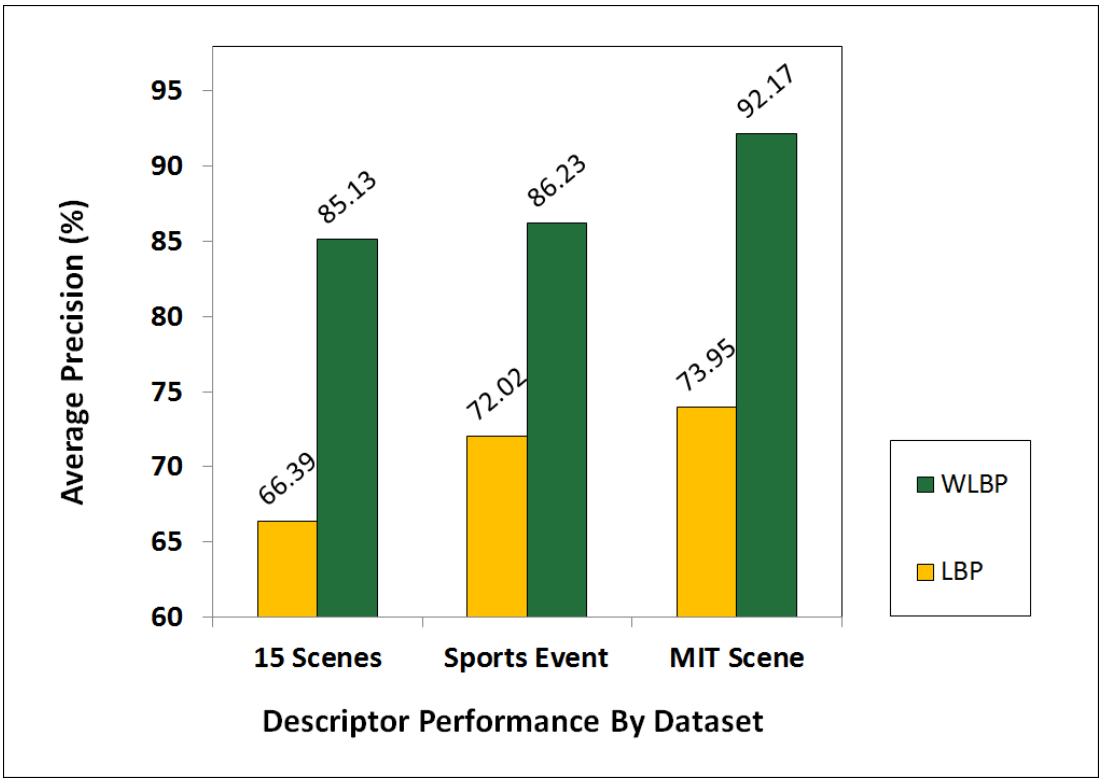


Figure 7.9 Comparison of the classification performance of the LBP and the proposed WLBP descriptors using an SVM classifier with a Hellinger kernel on the three datasets.

For the MIT Scene dataset, the protocol defined in (Oliva and Torralba 2001) is adopted where 100 images from each class are used for training and the remaining images for testing the performance. Here also, the WLBP significantly improves over the LBP feature, by a margin of 18%, and achieves an average precision of 92.17% (as shown in Figure 7.9) which is a very good result for this dataset. Table 7.1 shows a comparative evaluation of results obtained by other methods and by the proposed descriptor on this dataset.

On the Fifteen Scene Categories dataset, 100 training images from each category are chosen and rest for testing and the results are measured from five runs of experiments. Here, the overall performance of the WLBP is 85.13% which is again, much better than the traditional LBP as is evident from Figure 7.9. In addition, the category-wise classification performances of the grayscale LBP and the proposed WLBP features is displayed in

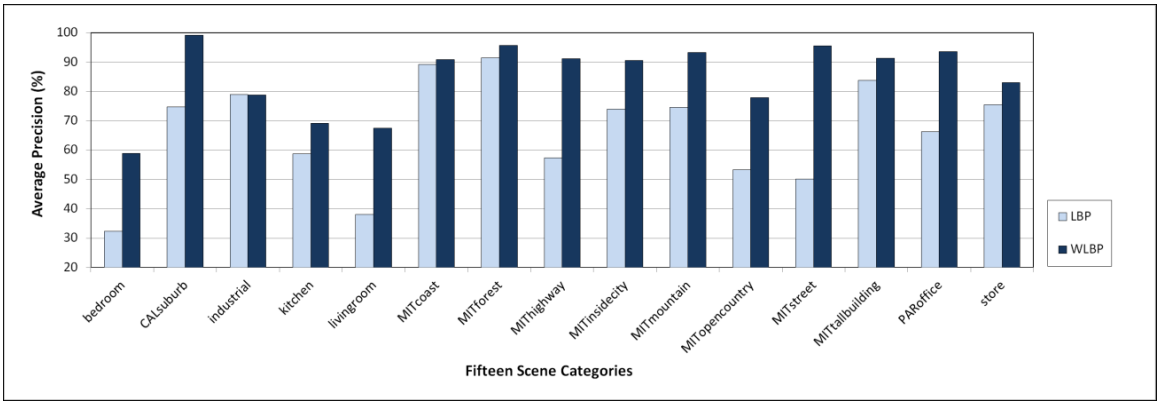


Figure 7.10 The comparative average classification performance of the LBP and the WLBP descriptors on the 15 categories of the Fifteen Scene Categories dataset.

Figure 7.10. Here, the horizontal axis reveals the fifteen scene categories, and the vertical axis displays the classification performance. A detailed comparison of the WLBP and other competitive methods on this dataset is given in Table 7.1.

7.3 Summary

In this chapter, a new local image descriptor has been presented for recognizing scene images by applying the Wigner distribution and a multi-neighborhood LBP technique on image patches. Combined with the DCT-based smoothing technique, the bag-of-visual words model and the spatial pyramid image representation and coupled with the SVM classifier, the new image descriptor significantly improves image classification performance over LBP. Experimental results on three popular scene image datasets show that the WLBP descriptor yields better classification performance than several recent state-of-the-art methods used by other researchers, such as the popular nonlinear Kernel Spatial Pyramid Matching (KSPM), SIFT Sparse-coded Spatial Pyramid Matching (ScSPM) and the Kernel Codebook (KC).

CHAPTER 8

CONCLUSIONS AND FUTURE WORK

This dissertation focuses on the feature extraction from grayscale and color images by proposing several new image descriptors based on wavelets, texture, color, shape and local features. The main contributions of this dissertation are as listed below:

- A new Gabor-based HOG method, the GHOG descriptor is proposed for grayscale and color images which enhances the HOG classification performance. The GHOG descriptor is analyzed in six different color spaces and the color GHOG features are then fused in an innovative way to produce the FC-GHOG descriptor that performs well in classifying different object and scene images. Experimental results using three grand challenge datasets, the Caltech 256 object categories dataset, the MIT Scene dataset, and the UIUC Sports Event dataset show that the proposed new image descriptors achieve better image classification performance than other popular image descriptors, such as the Scale Invariant Feature Transform (SIFT), the Pyramid Histograms of Oriented Gradients (PHOG), Spatial Envelope (SE), the Color SIFT four Concentric Circles (C4CC), Object Bank (OB), Context Aware Topic Model (CA-TM), as well as LBP.

- A new Gabor-PHOG (GPHOG) image descriptor is then created by enhancing the local features of an image using multiple Gabor filters for feature extraction. A comparative assessment of the classification performance of the GPHOG descriptor is made in grayscale and six different color spaces to further propose two novel color GPHOG descriptors. Finally, a Fused Color GPHOG (FC-GPHOG) descriptor is presented by integrating the Principal Component Analysis (PCA) features of the GPHOG descriptors in the six color spaces to combine color, shape and local feature information. Experimental results using two grand challenge datasets show that the

proposed new FC-GPHOG descriptor outperforms the PHOG and also achieves an image classification performance better than or comparable to other popular image descriptors, such as the Scale Invariant Feature Transform (SIFT) based Pyramid Histograms of visual Words (PHOW) descriptor, the Color SIFT four Concentric Circles (C4CC), Spatial Envelope, and Local Binary Patterns (LBP).

- Two new Gabor-based local, texture, shape and color feature extraction methods, namely the GLP and the FC-GLP are proposed that combines the GPHOG and the GLBP features using an optimal feature representation method such as PCA. The proposed descriptors exceed or achieve comparable performance to some of the best classification performances reported elsewhere. Experimental results carried out using three grand challenge datasets show that the FC-GLP descriptor improves classification performance over the GLBP and GPHOG descriptors and can be successfully applied for object and scene image classification.

- A novel Gabor-LBP-HOG (GLH) image descriptor is proposed which combines the GLBP and the GHOG descriptors to incorporate texture, shape and local information. The GLH descriptor is generated in six different color spaces as well as in grayscale. A new FC-GLH descriptor is also presented for object and scene image classification by integrating the GLH descriptors in the six different color spaces to further encode color information. Experimental results using three grand challenge datasets, the Caltech 256 object categories dataset, the MIT Scene dataset, and the UIUC Sports Event dataset show that the GLH and FC-GLH descriptors achieve better image classification performance than other popular image descriptors as well as GLBP, GHOG, GPHOG and GLP feature vectors.

- A new part based local image descriptor is presented for recognizing scene images by applying the Wigner distribution and a multi-neighborhood LBP technique on image patches. Combined with the DCT-based smoothing technique, the bag-of-

visual words model and the spatial pyramid image representation and coupled with the SVM classifier, the new image descriptor significantly improves image classification performance over LBP. Experimental results on three popular scene image datasets show that the WLBP descriptor yields better classification performance than several recent state-of-the-art methods used by other researchers, such as the popular nonlinear Kernel Spatial Pyramid Matching (KSPM), SIFT Sparse-coded Spatial Pyramid Matching (ScSPM) and the Kernel Codebook (KC).

Future work lies in the following directions:

- The proposed WLBP descriptor is found to be promising for image classification tasks. However, it is only created for grayscale images. One future direction of work would be to extend it to color images and calculate the WLBP for different color planes and investigate the image classification performance.

- The author would like to work on more datasets available for image search and classification.

- The descriptors proposed in this dissertation work on either whole image or small image patches taken from the entire image. One direction of future research would be to first detect regions of interest, and then develop a part-based image descriptor by extracting features from the selected object regions by combining the techniques discussed here with the part-based representation method.

- Using Gabor wavelets and Wigner distribution for designing the image descriptors achieved satisfactory classification results and enhanced performance. The author would like to further explore wavelets to construct new descriptors.

REFERENCES

R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

R. Arandjelović and A. Zisserman. All about VLAD. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

S. Banerji, A. Sinha, and C. Liu. New image descriptors based on color, texture, shape, and wavelets for object and scene image classification. *Neurocomputing*, 117(0): 173–185, 2013.

S. Banerji, A. Verma, and C. Liu. Novel color LBP descriptors for scene and image texture classification. In *15th International Conference on Image Processing, Computer Vision, and Pattern Recognition*, pages 537–543, Las Vegas, NV, USA, July 18-21 2011.

T. Barbu. Novel automatic video cut detection technique using gabor filtering. *Computers and Electrical Engineering*, 35(5):712–721, September 2009.

L. Bo, X. Ren, and D. Fox. Hierarchical matching pursuit for image classification: Architecture and fast algorithms. In *Neural Information Processing Systems*, pages 2115–2123, Granada, Spain, 2011.

A. Bosch, A. Zisserman, and X. Munoz. Scene classification via pLSA. In *The European Conference on Computer Vision*, pages 517–530, Graz, Austria, 2006.

A. Bosch, A. Zisserman, and X. Munoz. Image classification using random forests and ferns. In *The 11th International Conference on Computer Vision*, pages 1–8, Rio de Janeiro, Brazil, 2007a.

A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *International Conference on Image and Video Retrieval*, pages 401–408, Amsterdam, The Netherlands, July 9-11 2007b.

A. Bosch, A. Zisserman, and X. Munoz. Scene classification using a hybrid generative/discriminative approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(4):712–727, 2008.

M. Bratkova, S. Boulos, and P. Shirley. oRGB: A practical opponent color space for computer graphics. *IEEE Computer Graphics and Applications*, 29(1):42–55, 2009.

G. Burghouts and J.-M. Geusebroek. Performance evaluation of local color invariants. *Computer Vision and Image Understanding*, 113(1):48–62, 2009.

S. Chen and C. Liu. *Various Discriminatory Features for Eye Detection*. Cross Disciplinary Biometric Systems, C. Liu and V. Mago Eds., Springer, Berlin, Heidelberg, 2012.

W. Chen, M-J. Er, and S. Wu. Illumination compensation and normalization for robust face recognition using discrete cosine transform in logarithm domain. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 36(2):458–466, 2006.

M. Crosier and L.D. Griffin. Texture classification with a dictionary of basic image features. In *Computer Vision and Pattern Recognition*, pages 1–7, Anchorage, Alaska, June 23-28, 2008.

N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *The 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 886–893, San Diego, CA, USA, 2005.

J.G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional cortical filters. *Journal of the Optical Society of America*, 2(7):1160–1169, 1985.

J.G. Daugman. Complete discrete 2-d Gabor transforms by neural networks for image analysis and compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1169–1179, 1988.

J.G. Daugman. High confidence visual recognition of persons by a test of statistical independence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15 (11):1148–1161, 1993.

L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Conference on Computer Vision and Pattern Recognition*, pages 524–531, San Diego, CA, USA, 2005.

K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, second edition, 1990.

S. Gabarda and G. Cristóbal. Cloud covering denoising through image fusion. *Image and Vision Computing*, 25(5):523–530, 2007.

R.C. Gonzalez and R.E. Woods. *Digital Image Processing*. Prentice Hall, Upper Saddle River, NJ, 2001.

R.C. Gonzalez and R.E. Woods. *Digital Image Processing*. Pearson Prentice Hall, Upper Saddle River, NJ, third edition, 2008.

G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007. URL `http://authors.library.caltech.edu/7694`. Accessed: 04-21-2014.

J. Gu and C. Liu. Feature local binary patterns with application to eye detection. *Neurocomputing*, 113(0):138–152, 2013.

Z. M. Hafed and M. D. Levine. Face recognition using the discrete cosine transform. *International Journal of Computer Vision*, 43(3):167–188, July 2001.

L. Jacobson and H. Wechsler. Derivation of optical flow using a spatiotemporal-frequency approach. *Computer Vision, Graphics, and Image Processing*, 38(1):29–65, 1987.

A. K. Jain, S. Prabhakar, L. Hong, and S. Pankanti. Filterbank-based fingerprint matching. *IEEE Transactions on Image Processing*, 9(5):846–859, 2000.

J. Jones and L. Palmer. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, pages 1233–1258, 1987.

S. Lazebnik, C. Schmid, and J. Ponce. Semi-local affine parts for object recognition. In *British Machine Vision Conference*, volume 2, pages 959–968, London, September 7-9 2004.

S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2169–2178, New York, NY, USA, 2006.

H. Lee, Y. Chung, J. Kim, and D. Park. Face image retrieval using sparse representation classifier with gabor-lbp histogram. In *11th International Workshop on Information Security Applications*, pages 273–280, Jeju Island, Korea, 2010.

L.-J. Li and L. Fei-Fei. What, where and who? classifying event by scene and object recognition. In *IEEE International Conference in Computer Vision*, pages 1–8, 2007.

L-J. Li, H. Su, E. P. Xing, and L. Fei-Fei. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *Neural Information Processing Systems*, pages 1378–1386, Vancouver, Canada, 2010.

C. Liu. A Bayesian discriminating features method for face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(6):725–740, 2003.

C. Liu. Enhanced independent component analysis and its application to content based face image retrieval. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 34(2):1117–1127, 2004a.

C. Liu. Gabor-based kernel PCA with fractional power polynomial models for face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(5): 572–581, 2004b.

C. Liu. Capitalize on dimensionality increasing techniques for improving face recognition grand challenge performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(5):725–737, 2006.

C. Liu. The Bayes decision rule induced similarity measures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(29):1116–1117, 2007.

C. Liu. Learning the uncorrelated, independent, and discriminating color spaces for face recognition. *IEEE Transactions on Information Forensics and Security*, 3(2):213–222, 2008.

C. Liu. Extracting discriminative color features for face recognition. *Pattern Recognition Letters*, 32(14):1796–1804, 2011.

C. Liu and V. Mago, editors. *Cross Disciplinary Biometric Systems*. Springer, Berlin, Heidelberg, 2012.

C. Liu and H. Wechsler. Robust coding schemes for indexing and retrieval from large face databases. *IEEE Transactions on Image Processing*, 9(1):132–137, 2000.

C. Liu and H. Wechsler. Face recognition using independent Gabor wavelet features. In *3rd International Conference on Audio- and Video-Based Biometric Person Authentication*, Halmstad, Sweden, June 6-8, 2001.

C. Liu and H. Wechsler. Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition. *IEEE Transactions on Image Processing*, 11(4):467–476, 2002.

C. Liu and H. Wechsler. Independent component analysis of Gabor features for face recognition. *IEEE Transactions on Neural Networks*, 14(4):919–928, 2003.

C. Liu and J. Yang. ICA color space for pattern recognition. *IEEE Transactions on Neural Networks*, 2(20):248–257, 2009.

Z. Liu and C. Liu. Fusion of the complementary discrete cosine features in the YIQ color space for face recognition. *Computer Vision and Image Understanding*, 111(3): 249–262, 2008.

D.G. Lowe. Object recognition from local scale-invariant features. In *The International Conference on Computer Vision*, pages 1150–1157, Corfu, Greece, 1999.

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

O. Ludwig, D. Delgado, V. Goncalves, and U. Nunes. Trainable classifier-fusion schemes: An application to pedestrian detection. In *12th International IEEE Conference On Intelligent Transportation Systems*, volume 1, pages 432–437, St. Louis, USA, 2009.

B. S. Manjunath and W. Y. Ma. Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):837–842, 1996.

C. Mao, A. Gururajan, H. Sari-Sarraf, and E. F. Hequet. Machine vision scheme for stain-release evaluation using gabor filters with optimized coefficients. *Machine Vision and Applications*, 23(2):349–361, 2012.

S. Marcelja. Mathematical description of the responses of simple cortical cells. *Journal of the Optical Society of America*, 70:1297–1300, 1980.

K. Murphy, A. Torralba, and W. T. Freeman. Using the forest to see the trees: A graphical model relating features, objects, and scenes. In *Neural Information Processing Systems*. MIT Press, 2003.

W. Niblack, R. Barber, and W. Equitz. The QBIC project: Querying images by content using color, texture and shape. In *SPIE Conference on Geometric Methods in Computer Vision II*, pages 173–187, San Diego, CA, USA, 1993.

Z. Niu, G. Hua, X. Gao, and Q. Tian. Context aware topic model for scene recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2743–2750, Providence, RI, USA, June 16-21 2012.

E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In *9th European Conference on Computer Vision*, pages 490–503, Graz, Austria, 2006.

T. Ojala, M. Pietikainen, and D. Harwood. Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In *International Conference on Pattern Recognition*, pages 582–585, Jerusalem, Israel, 1994.

T. Ojala, M. Pietikainen, and D. Harwood. A comparative study of texture measures with classification based on feature distributions. *Pattern Recognition*, 29(1):51–59, 1996.

T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

M. Pontil and A. Verri. Support vector machines for 3D object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(6):637–646, 1998.

J. Sanchez, F. Perronnin, and T. Campos. Modeling the spatial layout of images beyond spatial pyramids. *Pattern Recognition Letters*, 33(16):2216–2223, 2012.

B. Schiele and J. Crowley. Recognition without correspondence using multidimensional receptive field histograms. *International Journal of Computer Vision*, 36(1):31–50, 2000.

P. Shih and C. Liu. Comparative assessment of content-based face image retrieval in different color spaces. *International Journal of Pattern Recognition and Artificial Intelligence*, 19(7):1039–1048, 2005.

A. Sinha, S. Banerji, and C. Liu. Novel color gabor-lbp-phog (glp) descriptors for object and scene image classification. In *The Eighth Indian Conference on Vision, Graphics and Image Processing*, page 58, Mumbai, India, December 16-19 2012.

J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. In *Ninth IEEE International Conference on Computer Vision*, pages 1470–1477, Nice, France, 2003.

A.R. Smith. Color gamut transform pairs. *Computer Graphics*, 12(3):12–19, 1978.

H. Stokman and T. Gevers. Selection and fusion of color models for image feature detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3): 371–381, 2007.

M.J. Swain and D.H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

A. Torralba, K. P. Murphy, W. T. Freeman, and Mark A. Rubin. Context-based vision system for place and object recognition. In *The Ninth IEEE International Conference on Computer Vision*, volume 1, pages 273–280, Nice, France, 2003.

V. G. Vaidya and R. M. Haralick. Wigner distribution for 2d motion estimation from noisy images. *Journal of Visual Communication and Image Representation*, 4(4):281–297, 1993.

J.C. Van Gemert, C.J. Veenman, A.W.M. Smeulders, and J.-M. Geusebroek. Visual word ambiguity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32 (7):1271–1283, 2010.

Y.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

A. Vedaldi and B. Fulkerson. Vlfeat — an open and portable library of computer vision algorithms. In *The 18th Annual ACM International Conference on Multimedia*, pages 1469–1472, Firenze, Italy, 2010.

A. Verma, S. Banerji, and C. Liu. A new color SIFT descriptor and methods for image category classification. In *International Congress on Computer Applications and Computational Science*, pages 819–822, Singapore, December 4-6 2010.

A. Verma and C. Liu. Novel EFM-KNN classifier and a new color descriptor for image classification. In *20th IEEE Wireless and Optical Communications Conference (Multimedia Services and Applications)*, Newark, NJ, USA, April 15-16 2011.

J. Ville. Theorie et Applications de la Notion de Signal Analytique. *Cables et Transmission*, 1:61–74, 1948.

E. Wigner. On the Quantum Correction For Thermodynamic Equilibrium. *Physical Review Online Archive (Prola)*, 40(5):749–759, 1932.

S. Xie, S. Shan, X. Chen, and J. Chen. Fusing local patterns of gabor magnitude and phase for face recognition. *IEEE Transactions on Image Processing*, 19(5):1349–1361, 2010.

J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1794–1801, Singapore, December 4-6 2009.

J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision*, 73(2):213–238, June 2007.

W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang. Local gabor binary pattern histogram sequence (lgbphs): A novel non-statistical model for face representation and recognition. In *The Tenth IEEE International Conference on Computer Vision*, pages 786–791, Beijing, China, 2005.

W. Zhao, Y. Jiang, and C. Ngo. Keyframe retrieval by keypoints: Can point-to-point matching help. In *The Fifth International Conference on Image and Video Retrieval*, pages 72–81, Tempe, AZ, USA, 2006.

C. Zhu, C. Bichot, and L. Chen. Multi-scale color local binary patterns for visual object classes recognition. In *International Conference on Pattern Recognition*, pages 3065–3068, Istanbul, Turkey, August 23-26 2010.