Chapter 7
SIFT Features in Multiple Color Spaces for Improved Image Classification

Abhishek Verma and Chengjun Liu

Abstract This chapter first discusses oRGB-SIFT descriptor, and then integrates it with other color SIFT features to produce the Color SIFT Fusion (CSF), the Color Grayscale SIFT Fusion (CGSF), and the CGSF+PHOG descriptors for image classification with special applications to image search and video retrieval. Classification is implemented using the EFM-NN classifier, which combines the Enhanced Fisher Model (EFM) and the Nearest Neighbor (NN) decision rule. The effectiveness of the proposed descriptors and classification method is evaluated using two large scale and challenging datasets: the Caltech 256 database and the UPOL Iris database. The experimental results show that (i) the proposed oRGB-SIFT descriptor improves recognition performance upon other color SIFT descriptors; and (ii) the CSF, the CGSF, and the CGSF+PHOG descriptors perform better than the other color SIFT descriptors.

7.1 Introduction

Content-based image retrieval is based on image similarity in terms of visual content such as features from color, texture, shape, etc. to a user-supplied query image or user-specified image features has been a focus of interest for the past several years. Color features provide powerful information for image search, indexing, and classification [26], [41], [32], in particular for identification of biometric images [38], [36], objects, natural scene, image texture and flower categories [39], [37], [2] and geographical features from images. The choice of a color space is important for many computer vision algorithms. Different color spaces display different color properties. With the large variety of available color spaces, the inevitable question

Abhishek Verma
California State University, Fullerton, CA 92834, USA, e-mail: averma@fullerton.edu

Chengjun Liu
New Jersey Institute of Technology, Newark, NJ 07102, USA, e-mail: chengjun.liu@njit.edu

that arises is how to select a color space that produces best results for a particular computer vision task. Two important criteria for color feature detectors are that they should be stable under varying viewing conditions, such as changes in illumination, shading, highlights, and they should have high discriminative power. Color features such as the color histogram, color texture and local invariant features provide varying degrees of success against image variations such as viewpoint and lighting changes, clutter and occlusions [9], [7], [34].

In the past, there has been much emphasis on the detection and recognition of locally affine invariant regions [27], [29]. Successful methods are based on representing a salient region of an image by way of an elliptical affine region, which describes local orientation and scale. After normalizing the local region to its canonical form, image descriptors are able to capture the invariant region appearance. Interest point detection methods and region descriptors can robustly detect regions, which are invariant to translation, rotation and scaling [27], [29]. Affine region detectors when combined with the intensity Scale-Invariant Feature Transform (SIFT) descriptor [27] has been shown to outperform many alternatives [29].

In this chapter, the SIFT descriptor is extended to different color spaces, including oRGB color space [6], oRGB-SIFT feature representation is proposed, furthermore it is integrated with other color SIFT features to produce the Color SIFT Fusion (CSF), and the Color Grayscale SIFT Fusion (CGSF) descriptors. Additionally, the CGSF is combined with the Pyramid of Histograms of Orientation Gradients (PHOG) to obtain the CGSF+PHOG descriptor for image category classification with special applications to biometrics. Classification is implemented using EFM-NN classifier [25], [24], which combines the Enhanced Fisher Model (EFM) and the Nearest Neighbor (NN) decision rule [12]. The effectiveness of the proposed descriptors and classification method is evaluated on two large scale, grand challenge datasets: the Caltech 256 dataset and the UPOL Iris database.

Rest of the chapter is organized as follows. In section 7.2 we review image-level global and local feature descriptors. Section 7.3 presents a review of five color spaces in which the color SIFT descriptors are defined followed by a discussion on clustering, visual vocabulary tree, and visual words for SIFT descriptors in section 7.4. Thereafter, in section 7.5 five conventional SIFT descriptors are presented: the RGB-SIFT, the rgb-SIFT, the HSV-SIFT, the YCbCr-SIFT, and the grayscale-SIFT descriptors and four new color SIFT descriptors are presented: the oRGB-SIFT, the Color SIFT Fusion (CSF), the Color Grayscale SIFT Fusion (CGSF), and the CGSF+PHOG descriptors. Section 7.6 presents a detailed discussion on the EFM-NN classification methodology. Description of datasets used for evaluation of methodology is provided in section 7.7. Next, in section 7.8 we present experimental results of evaluation of color SIFT descriptors. Section 7.9 concludes the chapter.

7.2 Related Work

In past years, use of color as a means to biometric image retrieval [26], [32], [22] and object and scene search has gained popularity. Color features can capture discriminative information by means of the color invariants, color histogram, color texture, etc. The earliest methods for object and scene classification were mainly based on the global descriptors such as the color and texture histogram [30], [31]. One of the earlier works is the color indexing system designed by Swain and Ballard, which uses the color histogram for image inquiry from a large image database [35]. Such methods are sensitive to viewpoint and lighting changes, clutter and occlusions. For this reason, global methods were gradually replaced by the part-based methods, which became one of the popular techniques in the object recognition community. Part-based models combine appearance descriptors from local features along with their spatial relationship. Harris interest point detector was used for local feature extraction; such features are only invariant to translation [1], [40]. Afterwards, local features with greater invariance were developed, which were found to be robust against scale changes [11] and affine deformations [20]. Learning and inference for spatial relations poses a challenging problem in terms of its complexity and computational cost. Whereas, the orderless bag-of-words methods [11], [21], [17] are simpler and computationally efficient, though they are not able to represent the geometric structure of the object or to distinguish between foreground and background features. For these reasons, the bag-of-words methods are not robust to clutter. One way to overcome this drawback is to design kernels that can yield high discriminative power in presence of noise and clutter [15].

Further, work on color based image classification appears in [26], [41], [23] that propose several new color spaces and methods for face classification and in [5] the HSV color space is used for the scene category recognition. Evaluation of local color invariant descriptors is performed in [7]. Fusion of color models, color region detection and color edge detection have been investigated for representation of color images [34]. Key contributions in color, texture, and shape abstraction have been discussed in Datta et al. [9].

As discussed before, many recent techniques for the description of images have considered local features. The most successful local image descriptor so far is Lowe's SIFT descriptor [27]. The SIFT descriptor encodes the distribution of Gaussian gradients within an image region. It is a 128-bin histogram that summarizes the local oriented gradients over 8 orientations and over 16 locations. This can efficiently represent the spatial intensity pattern, while being robust to small deformations and localization errors. Several modifications to the SIFT features have been proposed; among them are the PCA-SIFT [18], GLOH [28], and SURF [3]. These region-based descriptors have achieved a high degree of invariance to the overall illumination conditions for planar surfaces. Although, designed to retrieve identical object patches, SIFT-like features turn out to be quite successful in the bag-of-words approaches for general scene and object classification [5].

The Pyramid of Histograms of Orientation Gradients (PHOG) descriptor [4] is able to represent an image by its local shape and the spatial layout of the shape. The

local shape is captured by the distribution over edge orientations within a region, and the spatial layout 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.

7.3 Color Spaces

This section presents a review of five color spaces in which the color SIFT descriptors are defined.

7.3.1 RGB and rgb Color Spaces

A color image contains three component images, and each pixel of a color image is specified in a color space, which serves as a color coordinate system. The commonly used color space is the RGB color space. Other color spaces are usually calculated from the RGB color space by means of either linear or nonlinear transformations.

To reduce the sensitivity of the RGB images to luminance, surface orientation, and other photographic conditions, the rgb color space is defined by normalizing the R, G, and B components:

$$\begin{aligned} r &= R/(R+G+B) \\ g &= G/(R+G+B) \\ b &= B/(R+G+B) \end{aligned} \tag{7.1}$$

Due to the normalization r and g are scale-invariant and thereby invariant to light intensity changes, shadows and shading [13].

7.3.2 HSV Color Space

The HSV color space is motivated by human vision system because humans describe color by means of hue, saturation, and brightness. Hue and saturation define chrominance, while intensity or value specifies luminance [14]. The HSV color space is defined as follows [33]:

$$
\begin{aligned}
&Let \quad \begin{cases} MAX = max(R,G,B) \\ MIN = min(R,G,B) \\ \delta = MAX - MIN \end{cases} \\
&V = MAX \\
&S = \begin{cases} \frac{\delta}{MAX} & \text{if } MAX \neq 0 \\ 0 & \text{if } MAX = 0 \end{cases} \\
&H = \begin{cases} 60(\frac{G-B}{\delta}) & \text{if } MAX = R \\ 60(\frac{B-R}{\delta}+2) & \text{if } MAX = G \\ 60(\frac{R-G}{\delta}+4) & \text{if } MAX = B \\ not\ defined & \text{if } MAX = 0 \end{cases}
\end{aligned} \tag{7.2}
$$

7.3.3 YCbCr Color Space

The YCbCr color space is developed for digital video standard and television transmissions. In YCbCr, the RGB components are separated into luminance, chrominance blue, and chrominance red:

$$
\begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} + \begin{bmatrix} 65.4810 & 128.5530 & 24.9660 \\ -37.7745 & -74.1592 & 111.9337 \\ 111.9581 & -93.7509 & -18.2072 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{7.3}
$$

where the R,G,B values are scaled to $[0,1]$.

7.3.4 oRGB Color Space

The oRGB color space [6] has three channels L, $C1$ and $C2$. The primaries of this model are based on the three fundamental psychological opponent axes: white-black, red-green, and yellow-blue. The color information is contained in $C1$ and $C2$. The value of $C1$ lies within $[-1,1]$ and the value of $C2$ lies within $[-0.8660,0.8660]$. The L channel contains the luminance information and its values range between $[0,1]$:

$$
\begin{bmatrix} L \\ C1 \\ C2 \end{bmatrix} = \begin{bmatrix} 0.2990 & 0.5870 & 0.1140 \\ 0.5000 & 0.5000 & -1.0000 \\ 0.8660 & -0.8660 & 0.0000 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{7.4}
$$

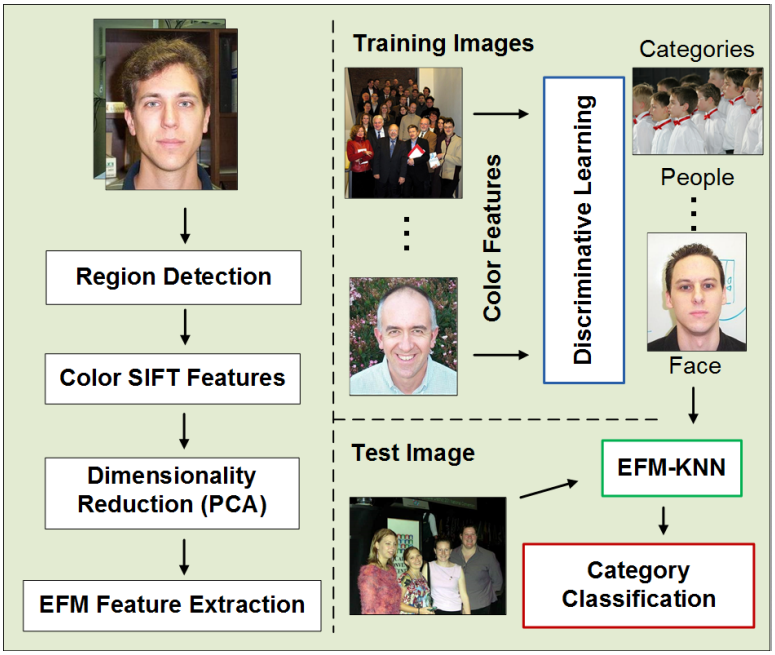

Fig. 7.1 An overview of SIFT feature extraction, learning and classification stages.

7.4 SIFT Feature Extraction, Clustering, Visual Vocabulary Tree, and Visual Words

This section first gives details of the SIFT feature extraction procedure. The next phase deals with the formation of visual vocabulary tree and visual words, here the normalized SIFT features are quantized with the vocabulary tree such that each image is represented as a collection of visual words, provided from a visual vocabulary. The visual vocabulary is obtained by vector quantization of descriptors computed from the training images using k-means clustering. See Figure 7.1 for an overview of the processing pipeline.

7.4.1 SIFT Feature Extraction

Image similarity may be defined in many ways based on the need of the application. It could be based on shape, texture, resolution, color or some other spatial features. The experiments here compute the SIFT descriptors extracted from the scale invariant points [42] on aforementioned color spaces. Such descriptors are called sparse descriptors, they have been previously used in [8], [19]. Scale invariant points are

obtained with the Hessian-affine point detector on the intensity channel. For the experiments, the Hessian-affine point detector is used because it has shown good performance in category recognition [29]. The remaining portion of feature extraction is then implemented according to the SIFT feature extraction pipeline of Lowe [27]. Canonical directions are found based on an orientation histogram formed on the image gradients. SIFT descriptors are then extracted relative to the canonical directions.

7.4.2 Clustering, Visual Vocabulary Tree, and Visual Words

The visual vocabulary tree defines a hierarchical quantization that is constructed with the hierarchical k-means clustering. A large set of representative descriptor vectors taken from the training images are used in the unsupervised training of the tree. Instead of k defining the final number of clusters or quantization cells, k defines the branch factor (number of children of each node) of the tree. First, an initial k-means process is run on the training data, defining k cluster centers. The training data is then partitioned into k groups, where each group consists of the descriptor vectors closest to a particular cluster center. The same process is then recursively applied to each group of descriptor vectors, recursively defining clusters by splitting each cluster into k new parts. The tree is determined level by level, up to some maximum number of levels say L, and each division into k parts is only defined by the distribution of the descriptor vectors that belong to the parent cluster. Once the tree is computed, its leaf nodes are used for quantization of descriptors from the training and test images.

It has been experimentally observed that most important for the retrieval quality is to have a large vocabulary, i.e., large number of leaf nodes. While the computational cost of increasing the size of the vocabulary in a non-hierarchical manner would be very high, the computational cost in the hierarchical approach is logarithmic in the number of leaf nodes. The memory usage is linear in the number of leaf nodes kL. The current implementation builds a tree of $6,561$ leaf nodes and $k = 9$. See Figure 7.2 for an overview of the quantization process.

To obtain fixed-length feature vectors per image, the visual words model is used [5], [8]. The visual words model performs vector quantization of the color descriptors in an image against a visual vocabulary. In the quantization phase, each descriptor vector is simply propagated down the tree at each level by comparing the descriptor vector to the k candidate cluster centers (represented by k children in the tree) and choosing the closest one till it is assigned to a particular leaf node. This is a simple matter of performing k dot products at each level, resulting in a total of kL dot products, which is very efficient if k is not too large.

Once all the SIFT features from an image are quantized, a fixed length feature vector would be obtained. The feature vector is normalized to zero mean and unit standard deviation. The advantage of representing an image as a fixed length feature vector lies in the fact that it allows to effectively compare images that vary in size.

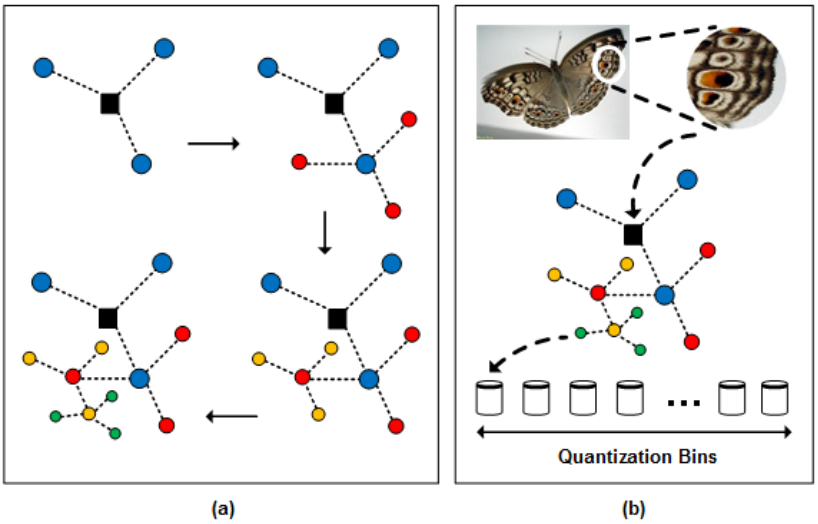

Fig. 7.2 (a) An illustration of the process of constructing a vocabulary tree by hierarchical k-means. The hierarchical quantization is defined at each level by k centers (in this case $k = 3$). (b) A large number of elliptical regions are extracted from the image and normalized to circular regions. A SIFT descriptor vector is computed for each region. The descriptor vector is then hierarchically quantized by the vocabulary tree. The number of quantization bins is the number of leaf nodes in the vocabulary tree; this is the length of the final feature vector as well.

7.5 Color SIFT Descriptors

The SIFT descriptor proposed by Lowe transforms an image into a large collection of feature vectors, each of which is invariant to image translation, scaling, and rotation, partially invariant to the illumination changes, and robust to local geometric distortion [27]. The key locations used to specify the SIFT descriptor are defined as maxima and minima of the result of the difference of Gaussian function applied in the scale-space to a series of smoothed and resampled images. SIFT descriptors robust to local affine distortions are then obtained by considering pixels around a radius of the key location.

The grayscale SIFT descriptor is defined as the SIFT descriptor applied to the grayscale image. A color SIFT descriptor in a given color space is derived by individually computing the SIFT descriptor on each of the three component images in the specific color space. This produces a 384 dimensional descriptor that is formed from concatenating the 128 dimensional vectors from the three channels. As a result, four conventional color SIFT descriptors are defined: the RGB-SIFT, the YCbCr-SIFT, the HSV-SIFT, and the rgb-SIFT descriptors.

Furthermore, four new color SIFT descriptors are defined in the oRGB color space and the fusion in different color spaces. In particular, the oRGB-SIFT descriptor is constructed by concatenating the SIFT descriptors of the three compo-

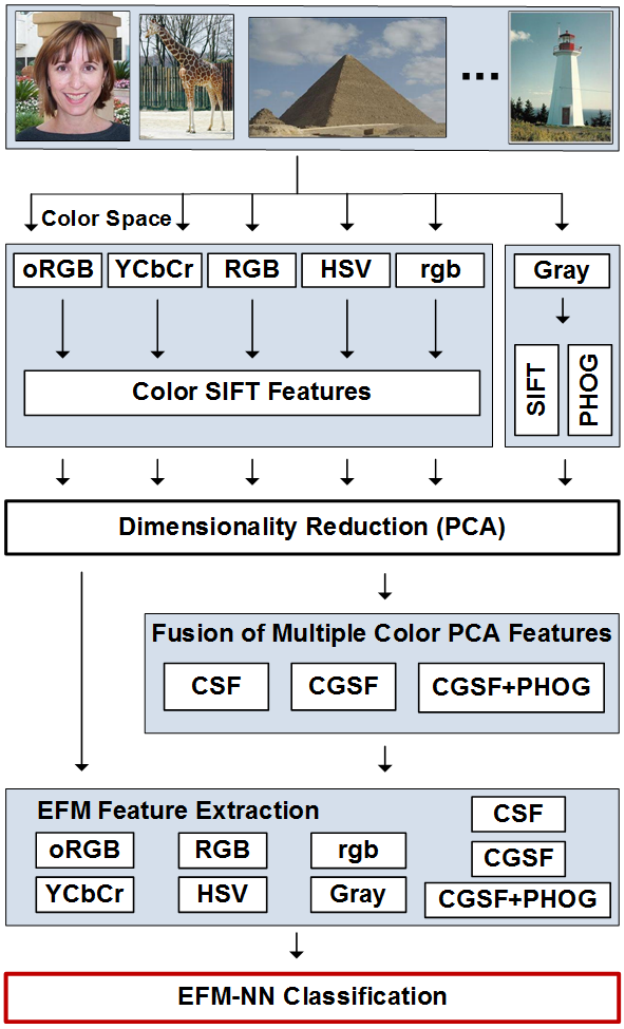

Fig. 7.3 Multiple Color SIFT features fusion methodology using the EFM feature extraction.

nent images in the oRGB color space. The Color SIFT Fusion (CSF) descriptor is formed by fusing the RGB-SIFT, the YCbCr-SIFT, the HSV-SIFT, the oRGB-SIFT, and the rgb-SIFT descriptors. The Color Grayscale SIFT Fusion (CGSF) descriptor is obtained by fusing further the CSF descriptor and the grayscale-SIFT descriptor. The CGSF is combined with the Pyramid of Histograms of Orientation Gradients (PHOG) descriptor to obtain the CGSF+PHOG descriptor. See Figure 7.3 for multiple Color SIFT features fusion methodology.

7.6 EFM-NN Classifier

Image classification using the descriptors introduced in the preceding section is implemented using EFM-NN classifier [25], [24], which combines the Enhanced Fisher Model (EFM) and Nearest Neighbor (NN) decision rule [12]. Let $\mathscr{X} \in \mathbb{R}^N$ be a random vector whose covariance matrix is $\Sigma_{\mathscr{X}}$:

$$\Sigma_{\mathscr{X}} = \mathscr{E}\{[\mathscr{X} - \mathscr{E}(\mathscr{X})][\mathscr{X} - \mathscr{E}(\mathscr{X})]^t\} \tag{7.5}$$

where $\mathscr{E}(\cdot)$ is the expectation operator and t denotes the transpose operation. The eigenvectors of the covariance matrix $\Sigma_{\mathscr{X}}$ can be derived by PCA:

$$\Sigma_{\mathscr{X}} = \Phi \Lambda \Phi^t \tag{7.6}$$

where $\Phi = [\phi_1 \phi_2 \ldots \phi_N]$ 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. An important application of PCA is dimensionality reduction:

$$\mathscr{Y} = P^t \mathscr{X} \tag{7.7}$$

where $P = [\phi_1 \phi_2 \ldots \phi_K]$, and $K < N$. $\mathscr{Y} \in \mathbb{R}^K$ thus is composed of the most significant principal components. PCA, which is derived based on an optimal representation criterion, usually does not lead to good image classification performance. To improve upon PCA, the Fisher Linear Discriminant (FLD) analysis [12] is introduced to extract the most discriminating features.

The FLD method optimizes a criterion defined on the within-class and between-class scatter matrices, S_w and S_b [12]:

$$S_w = \sum_{i=1}^{L} P(\omega_i) \mathscr{E}\{(\mathscr{Y} - M_i)(\mathscr{Y} - M_i)^t | \omega_i\} \tag{7.8}$$

$$S_b = \sum_{i=1}^{L} P(\omega_i)(M_i - M)(M_i - M)^t \tag{7.9}$$

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. The criterion the FLD method optimizes is $J_1 = tr(S_w^{-1} S_b)$, which is maximized when Ψ contains the eigenvectors of the matrix $S_w^{-1} S_b$ [12]:

$$S_w^{-1} S_b \Psi = \Psi \Delta \tag{7.10}$$

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

$$\mathscr{Z} = \Psi^t \mathscr{Y} \tag{7.11}$$

$\mathscr{Z}$ thus is more effective than the feature vector $\mathscr{Y}$ derived by PCA 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 elimi-

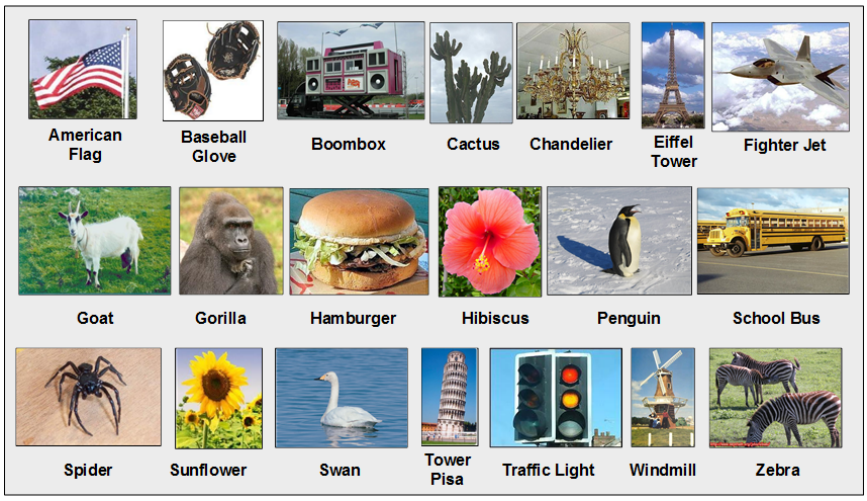

Fig. 7.4 Example images from the Caltech 256 object categories dataset.

nating the small-valued trailing eigenvalues of the within-class scatter matrix [24]. Enhanced Fisher Model (EFM), is capable of improving the generalization performance of the FLD method [24]. 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 [24]. The simultaneous diagonalization is stepwise equivalent to two operations as pointed out by [12]: whitening the within-class scatter matrix and applying PCA to the between-class scatter matrix using the transformed data. The stepwise operation shows that during whitening the eigenvalues of the within-class scatter matrix appear in the denominator. Since the small (trailing) eigenvalues tend to capture noise [24], they cause the whitening step to fit for misleading variations, which leads to poor generalization performance. To achieve enhanced 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) [24].

Image classification is implemented with EFM-NN, which uses nearest neighbor and cosine distance measure. Figure 7.3 shows the fusion methodology of multiple descriptors using EFM feature extraction and EFM-NN classification.

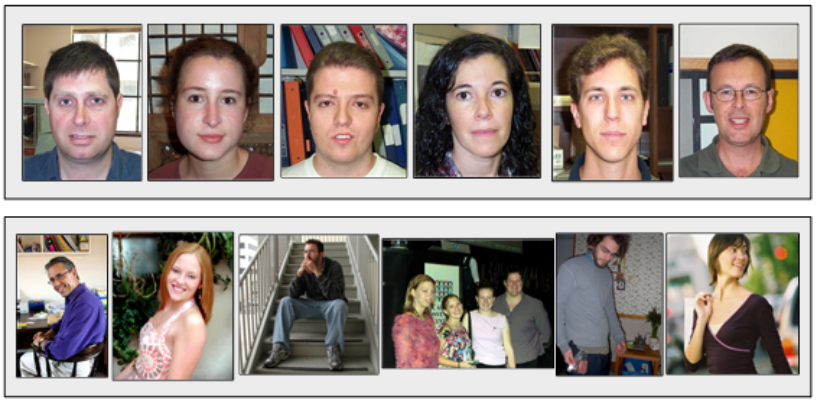

Fig. 7.5 Example images from the Faces and People classes of the Caltech 256 object categories dataset.

7.7 Description of Dataset

We perform experimental evaluation of Color SIFT descriptors on two publicly available large scale grand challenge datasets: the Caltech 256 object categories dataset and the UPOL iris dataset.

7.7.1 Caltech 256 Object Categories Dataset

The Caltech 256 dataset [16] comprises of 30,607 images divided into 256 categories and a clutter class. See Figure 7.4 for some images from the object categories and Figure 7.5 for some sample images from the Faces and People categories. The images have high intra-class variability and high object location variability. Each category contains at least 80 images, a maximum of 827 images and the mean number of images per category is 119. The images have been collected from Google and PicSearch, they represent a diverse set of lighting conditions, poses, back-grounds, image sizes, and camera systematics. The various categories represent a wide variety of natural and artificial objects in various settings. The images are in color, in JPEG format with only a small number of grayscale images. The average size of each image is 351x351 pixels.

7.7.2 UPOL Iris Dataset

The UPOL iris dataset [10] contains 128 unique eyes (or classes) belonging to 64 subjects with each class containing three sample images. The images of the left and right eyes of a person belong to different classes. The irises were scanned by a TOPCON TRC50IA optical device connected with a SONY DXC-950P 3CCD camera. The iris images are in 24-bit PNG format (color) and the size of each image is 576x768 pixels. See Figure 7.6 for some sample images from this dataset.

7.8 Experimental Evaluation of Color SIFT Descriptors on the Caltech 256 and the UPOL Iris Datasets

7.8.1 Experimental Methodology

In order to make a comparative assessment of the descriptors and methods; from the aforementioned two datasets we the Biometric 100 dataset with 100 categories includes the Iris category from the UPOL dataset, Faces and People categories and 97 randomly chosen categories from the Caltech 256 dataset. This dataset is of high difficulty due to the large number of classes with high intra-class and low inter-class variations.

The classification task is to assign each test image to one of a number of categories. The performance is measured using a confusion matrix, and the overall performance rates are measured by the average value of the diagonal entries of the confusion matrix. Dataset is split randomly into two separate sets of images for training and testing. From each class 60 images for training and 20 images for testing are randomly selected. There is no overlap in the images selected for training and testing. The classification scheme on the dataset compares the overall and category wise performance of ten different descriptors: the oRGB-SIFT, the YCbCr-SIFT, the RGB-SIFT, the HSV-SIFT, the rgb-SIFT, the PHOG, the grayscale-SIFT, the CSF, the CGSF, and the CGSF+PHOG descriptors. Classification is implemented using EFM-NN classifier, which combines the Enhanced Fisher Model (EFM) and the Nearest Neighbor (NN) decision rule.

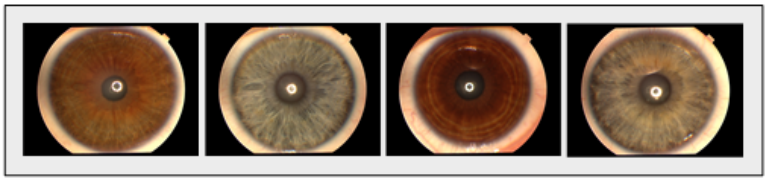

Fig. 7.6 Example images from the UPOL Iris dataset.

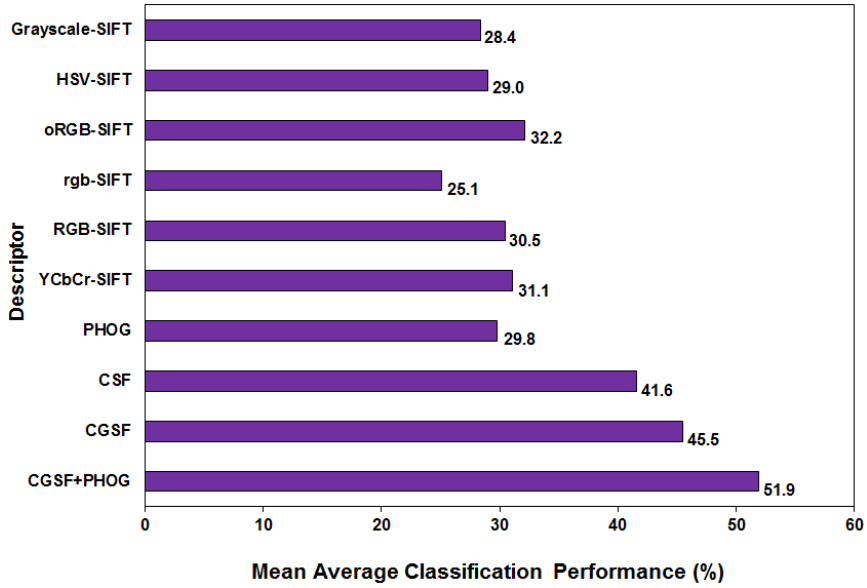

Fig. 7.7 The mean average classification performance of the ten descriptors: the oRGB-SIFT, the YCbCr-SIFT, the RGB-SIFT, the HSV-SIFT, the rgb-SIFT, the grayscale-SIFT, the PHOG, the CSF, the CGSF, and the CGSF+PHOG descriptors on the Biometric 100 dataset.

Table 7.1 Comparison of Classifiers across Ten Descriptors (%) on the Biometric 100 Dataset

Descriptor	PCA	EFM-NN
RGB-SIFT	27.9	**30.5**
HSV-SIFT	26.1	**29.0**
rgb-SIFT	23.1	**25.1**
oRGB-SIFT	29.4	**32.2**
YCbCr-SIFT	28.2	**31.1**
SIFT	26.3	**28.4**
PHOG	28.0	**29.8**
CSF	40.2	**41.6**
CGSF	44.6	**45.5**
CGSF+PHOG	49.4	**51.9**

7.8.2 Experimental Results on the Biometric 100 Categories Dataset

7.8.2.1 Evaluation of Overall Classification Performance of Descriptors with the EFM-NN Classifier

The first set of experiments assesses the overall classification performance of the ten descriptors on the Biometric 100 dataset with 100 categories. Note that for each

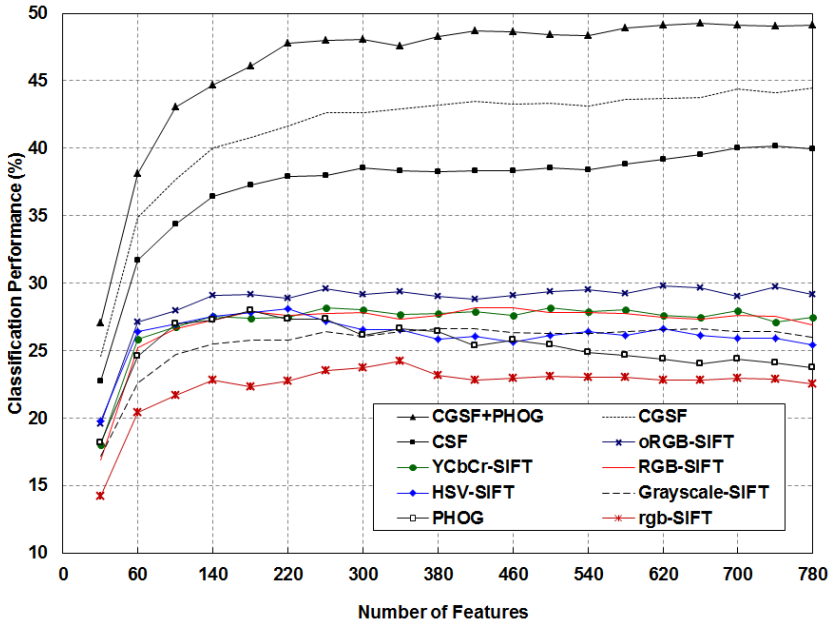

Fig. 7.8 Classification results using the PCA method across the ten descriptors with varying number of features on the Biometric 100 dataset.

category a five-fold cross validation is implemented for each descriptor using the EFM-NN classification technique to derive the average classification performance. As a result, each descriptor yields 100 average classification rates corresponding to the 100 image categories. The mean value of these 100 average classification rates is defined as the mean average classification performance for the descriptor.

The best recognition rate that is obtained is 51.9% from the CGSF+PHOG, which is a very respectable value for a dataset of this size and complexity. The oRGB-SIFT achieves the classification rate of 32.2% and hence once again outperforms other color descriptors. The success rate for YCbCr-SIFT comes in second place with 31.1% followed by the RGB-SIFT at 30.5%. Fusion of color SIFT descriptors (CSF) improves over the grayscale-SIFT by a huge 13.2%. Again, the grayscale-SIFT shows more distinctiveness than the rgb-SIFT, and improves the fusion (CGSF) result by a good 3.9% over the CSF. Fusing the CGSF and PHOG further improves the recognition rate over the CGSF by 6.4%. See Figure 7.7 for mean average classification performance of various descriptors.

7.8.2.2 Comparison of PCA and EFM-NN Results

The second set experiments compares the classification performance of the PCA and the EFM-NN (nearest neighbor) classifiers. Table 7.1 shows the results of the

two classifiers across various descriptors. It can be seen that the EFM-NN technique improves over the PCA technique by 2% to 3% on the color SIFT descriptors, by 2.1% on the grayscale-SIFT, and by 1.9% on the PHOG. The improvement on fused descriptors is in the range of 1%-2.6%. These results reaffirm the superiority of the EFM-NN classifier over the PCA technique.

7.8.2.3 Evaluation of PCA and EFM-NN Results upon Varying Number of Features

The third set of experiments evaluates the classification performance using the PCA and the EFM-NN methods respectively by varying the number of features over the following ten descriptors: CGSF+PHOG, CGSF, CSF, YCbCr-SIFT, oRGB-SIFT, RGB-SIFT, HSV-SIFT, Grayscale-SIFT, rgb-SIFT, and PHOG.

Classification performance is computed for up to 780 features with the PCA classifier. From Figure 7.8 it can be seen that the success rate for the CGSF+PHOG stays consistently above that of the CGSF and CSF over varying number of features and peaks at around 660 features. These three descriptors show an increasing trend overall and flatten out toward the end. The oRGB-SIFT, YCbCr-SIFT, RGB-SIFT, and grayscale-SIFT show a similar increasing trend and flatten toward the end. The oRGB-SIFT descriptor consistently stays above other color SIFT descriptors. The HSV-SIFT and PHOG peak in the first half of the graph and show a declining trend thereafter. The grayscale-SIFT maintains its superior performance upon the rgb-SIFT on the varying number of features.

With the EFM-NN classifier, the success rates are computed for up to 95 features. From Figure 7.9 it can be seen that the success rate for the CGSF+PHOG stays consistently above that of the CGSF and CSF over varying number of features and peaks at about 80 features. These three descriptors show an increasing trend throughout and tend to flatten above 65 features. The oRGB-SIFT consistently stays above the rest of the descriptors. The grayscale-SIFT improves over the rgb-SIFT but falls below the PHOG.

7.8.2.4 Evaluation of Descriptors and Classifier on Individual Image Categories

The fourth set of experiments assesses the eight descriptors using the EFM-NN classifier on individual image categories. Here a detailed analysis of the performance of the descriptors is performed with the EFM-NN classifier over 100 image categories. First the classification results on the three biometric categories are presented. From Table 7.2 it can be seen that the Iris category has a 100% recognition rate across all the descriptors. For the Faces category the color SIFT descriptors outperform the grayscale-SIFT by 5% to 10% and the fusion of all descriptors (CGSF+PHOG) reaches a 95% success rate. The People category achieves a high success rate of 40% with the CGSF+PHOG, surprisingly grayscale-SIFT outperforms the color descrip-

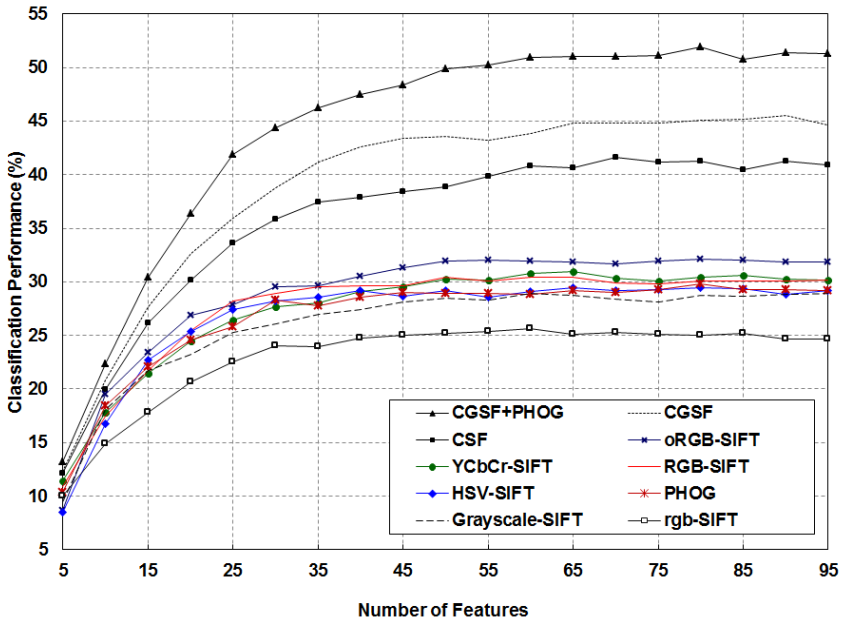

Fig. 7.9 Classification results using the EFM-NN method across the ten descriptors with varying number of features on the Biometric 100 dataset.

tors by 10% to 20%. The fusion of individual SIFT descriptors (CGSF) improves the classification performance for the People category.

The average success rate for the CGSF+PHOG over the top 20 categories is 90% with ten categories above the 90% mark. Individual color SIFT features improve upon the grayscale-SIFT on most of the categories, in particular for the Swiss army knife, Watch, American flag, and Roulette wheel categories. The CSF almost always improves over the grayscale-SIFT, with the exception of People and French horn categories. The CGSF either is at par or improves over the CSF for all categories with the exception of two of the categories. Most categories perform at their best when the PHOG is combined with the CGSF.

7.8.2.5 Evaluation of Descriptors and Classifier Based on Correctly Recognized Images

The final set of experiments further assesses the performance of the descriptors based on the correctly recognized images. See Figure 7.10(a) for some examples of the correctly classified images from the Iris, Faces, and People categories. Once again notice the high intra-class variability in the recognized images for the Faces and People class. Figure 7.10(b) shows some images from the Faces and People categories that are not recognized by the grayscale-SIFT but are correctly recog-

Table 7.2 Category Wise Descriptor Performance (%) Split-out with the EFM-NN Classifier on the Biometric 100 Dataset (Note That the Categories are Sorted on the CGSF+PHOG Results)

Category	CGSF+ PHOG	CGSF	CSF	oRGB SIFT	YCbCr SIFT	RGB SIFT	Gray SIFT	PHOG
iris	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
faces	**95**	90	90	90	**95**	90	85	**95**
people	**40**	**40**	25	20	20	15	30	10
hibiscus	**100**	**100**	95	70	80	85	75	55
french horn	**95**	85	85	85	65	80	90	20
leopards	95	90	**100**	90	95	95	**100**	90
saturn	**95**	**95**	**95**	**95**	85	90	**95**	55
school bus	**95**	**95**	**95**	75	85	**95**	80	60
swiss army knife	**95**	90	80	65	75	65	65	25
watch	**95**	60	55	45	40	45	30	85
zebra	**95**	80	60	60	35	40	45	60
galaxy	**90**	85	85	85	70	65	80	15
american flag	**85**	**85**	80	55	75	65	40	5
cartman	**85**	75	75	40	55	65	55	30
desk-globe	**85**	75	75	60	65	65	45	80
harpsichord	**85**	80	**85**	50	80	70	60	55
ketch	**85**	**85**	**85**	45	50	45	50	70
roulette wheel	**85**	80	75	70	65	75	55	35
hawksbill	**80**	**80**	75	55	60	70	55	40
iris flower	**80**	75	75	35	65	**80**	65	30
mountain bike	80	85	**90**	70	65	85	75	70

nized by the oRGB-SIFT. Figure 7.10(c) shows some images that are not recognized by the oRGB-SIFT but are correctly recognized by the CSF. Figure 7.10(d) shows some images from the People class, which are not recognized by the CSF but are correctly recognized by the CGSF+PHOG descriptor. Thus, combining grayscale-SIFT, PHOG, and CSF lends more discriminative power. Lastly in Figure 7.10(e) a face image unrecognized by the PCA but recognized by the EFM-NN classifier on the CGSF+PHOG descriptor.

See Figure 7.11(a) for some examples of the images unrecognized by the grayscale-SIFT but are correctly recognized by the oRGB-SIFT. Figure 7.11(b) shows some images that are not recognized by the oRGB-SIFT but are correctly recognized by the CSF. Figure 7.11(c) shows some images unrecognized by the CSF but are correctly recognized by the CGSF+PHOG descriptor. Lastly in Figure 7.11(d) images unrecognized by the PCA but recognized by the EFM-NN classifier on the CGSF+PHOG descriptor.

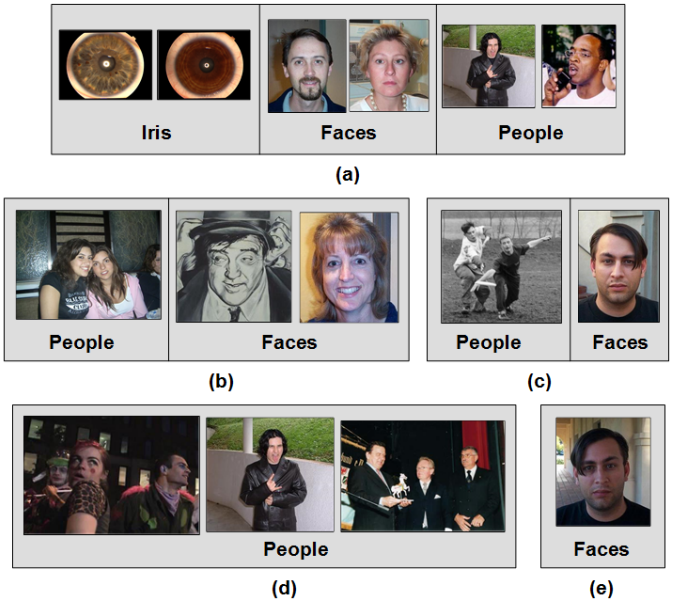

Fig. 7.10 Image recognition using the EFM-NN classifier on the Biometric 100 dataset: (a) examples of the correctly classified images from the three biometric image categories; (b) images unrecognized using the grayscale-SIFT descriptor but recognized using the oRGB-SIFT descriptor; (c) images unrecognized using the oRGB-SIFT descriptor but recognized using the CSF descriptor; (d) images unrecognized using the CSF but recognized using the CGSF+PHOG; (e) images unrecognized by PCA but recognized by EFM-NN on the CGSF+PHOG descriptor.

7.9 Conclusion

In this chapter we presented the oRGB-SIFT feature descriptor and its integration with other color SIFT features to produce the Color SIFT Fusion (CSF), the Color Grayscale SIFT Fusion (CGSF), and the CGSF+PHOG descriptors. Experimental results using two large scale and challenging datasets show that our oRGB-SIFT descriptor improves the recognition performance upon other color SIFT descriptors, and the CSF, the CGSF, and the CGSF+PHOG descriptors perform better than the other color SIFT descriptors. The fusion of the Color SIFT descriptors (CSF) and the Color Grayscale SIFT descriptor (CGSF) show significant improvement in the classification performance, which indicates that the various color-SIFT descriptors and the grayscale-SIFT descriptor are not redundant for image classification.

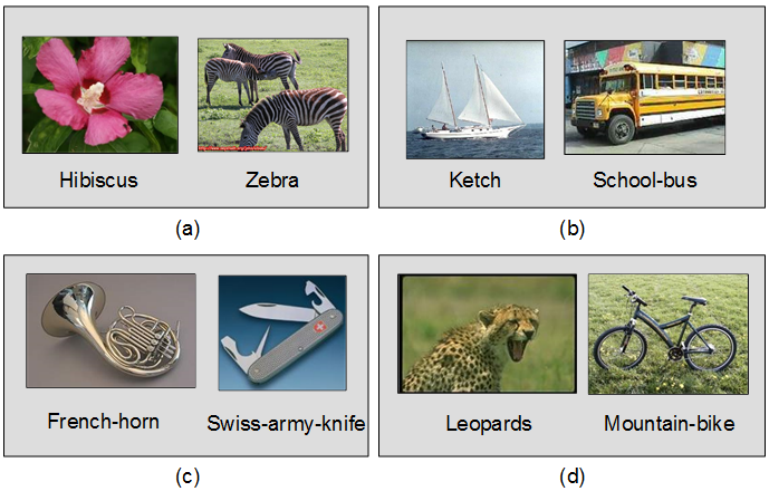

Fig. 7.11 Image recognition using the EFM-NN classifier on the Biometric 100 dataset: (a) example images unrecognized using the grayscale-SIFT descriptor but recognized using the oRGB-SIFT descriptor; (b) example images unrecognized using the oRGB-SIFT descriptor but recognized using the CSF descriptor; (c) images unrecognized using the CSF but recognized using the CGSF+PHOG. (d) Images unrecognized using the PCA but recognized using the EFM-NN on the CGSF+PHOG descriptor.

References

1. Agarwal, S., Roth, D.: Learning a sparse representation for object detection. In: European Conference on Computer Vision, vol. 4, pp. 113–130. Copenhagen, Denmark (2002)
2. Banerji, S., A., V., Liu, C.: Novel color LBP descriptors for scene and image texture classification. In: 15th Intl. Conf. on Image Processing, Computer Vision, and Pattern Recognition. Las Vegas, Nevada (2011)
3. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded up robust features. Computer Vision and Image Understanding **110**(3), 346–359 (2008)
4. Bosch, A., Zisserman, A., Munoz, X.: Representing shape with a spatial pyramid kernel. In: Int. Conf. on Image and Video Retrieval, pp. 401–408. Amsterdam, The Netherlands (2007)
5. Bosch, A., Zisserman, A., Munoz, X.: Scene classification using a hybrid generative/discriminative approach. IEEE Trans. on Pattern Analysis and Machine Intelligence **30**(4), 712–727 (2008)
6. Bratkova, M., Boulos, S., Shirley, P.: oRGB: A practical opponent color space for computer graphics. IEEE Computer Graphics and Applications **29**(1), 42–55 (2009)
7. Burghouts, G., Geusebroek, J.M.: Performance evaluation of local color invariants. Computer Vision and Image Understanding **113**, 48–62 (2009)
8. Csurka, G., Bray, C., Dance, C., Fan, L.: Visual categorization with bags of keypoints. In: Proc. Workshop Statistical Learning in Computer Vision, pp. 1–22 (2004)
9. Datta, R., Joshi, D., Li, J., Wang, J.: Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys **40**(2), 509–522 (2008)
10. Dobes, M., Martinek, J., Skoupil, D., Dobesova, Z., Pospisil, J.: Human eye localization using the modified Hough transform. Optik **117**(10), 468–473 (2006)

11. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: IEEE Conf. on Computer Vision and Pattern Recognition, vol. 2, pp. 264–271. Madison, Wisconsin (2003)
12. Fukunaga, K.: Introduction to Statistical Pattern Recognition, 2nd edn. Academic Press, San Diego, California, USA (1990)
13. Gevers, T., van de Weijer, J., Stokman, H.: Color feature detection: An overview. In: R. Lukac, K. Plataniotis (eds.) Color Image Processing: Methods and Applications. CRC Press, University of Toronto, Ontario, Canada (2006)
14. Gonzalez, C., Woods, R.: Digital Image Processing. Prentice Hall, Upper Saddle River, NJ, USA (2001)
15. Grauman, K., Darrell, T.: Pyramid match kernels: Discriminative classification with sets of image features. In: Int. Conference on Computer Vision, vol. 2, pp. 1458–1465. Beijing (2005)
16. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset. Tech. rep., California Institute of Technology (2007)
17. Jurie, F., Triggs, B.: Creating efficient codebooks for visual recognition. In: Int. Conference on Computer Vision, pp. 604–610. Beijing (2005)
18. Ke, Y., Sukthankar, R.: PCA-SIFT: A more distinctive representation for local image descriptors. In: IEEE Conf. on Computer Vision and Pattern Recognition, vol. 2, pp. 506–513. Washington, D.C. (2004)
19. Lazebnik, S., Schmid, C., Ponce, J.: A sparse texture representation using affine-invariant regions. In: IEEE Conf. on Computer Vision and Pattern Recognition, vol. 2, pp. 319–324. Madison, Wisconsin (2003)
20. Lazebnik, S., Schmid, C., Ponce, J.: Semi-local affine parts for object recognition. In: British Machine Vision Conference, vol. 2, pp. 959–968. London (2004)
21. Leung, T., Malik, J.: Representing and recognizing the visual appearance of materials using three-dimensional textons. Int. Journal of Computer Vision **43**(1), 29–44 (2001)
22. Liu, C.: Capitalize on dimensionality increasing techniques for improving face recognition grand challenge performance. IEEE Trans. Pattern Analysis and Machine Intelligence **28**(5), 725–737 (2006)
23. Liu, C.: Learning the uncorrelated, independent, and discriminating color spaces for face recognition. IEEE Trans. on Information Forensics and Security **3**(2), 213–222 (2008)
24. Liu, C., Wechsler, H.: Robust coding schemes for indexing and retrieval from large face databases. IEEE Trans. on Image Processing **9**(1), 132–137 (2000)
25. Liu, C., Wechsler, H.: Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition. IEEE Trans. on Image Processing **11**(4), 467–476 (2002)
26. Liu, C., Yang, J.: ICA color space for pattern recognition. IEEE Trans. on Neural Networks **2**(20), 248–257 (2009)
27. Lowe, D.: Distinctive image features from scale-invariant keypoints. Int. Journal of Computer Vision **60**(2), 91–110 (2004)
28. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Trans. on Pattern Analysis and Machine Intelligence **27**(10), 1615–1630 (2005)
29. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Van Gool, L.: A comparison of affine region detectors. Int. Journal of Computer Vision **65**(1-2), 43–72 (2005)
30. Pontil, M., Verri, A.: Support vector machines for 3D object recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence **20**(6), 637–646 (1998)
31. Schiele, B., Crowley, J.: Recognition without correspondence using multidimensional receptive field histograms. Int. Journal of Computer Vision **36**(1), 31–50 (2000)
32. Shih, P., Liu, C.: Comparative assessment of content-based face image retrieval in different color spaces. Int. Journal of Pattern Recognition and Artificial Intelligence **19**(7), 873–893 (2005)
33. Smith, A.: Color gamut transform pairs. Computer Graphics **12**(3), 12–19 (1978)
34. Stokman, H., Gevers, T.: Selection and fusion of color models for image feature detection. IEEE Trans. on Pattern Analysis and Machine Intelligence **29**(3), 371–381 (2007)

35. Swain, M., Ballard, D.: Color indexing. Int. Journal of Computer Vision **7**(1), 11–32 (1991)
36. Verma, A., Liu, C.: Fusion of color SIFT features for image classification with applications to biometrics. In: 11th IAPR International Conference on Pattern Recognition and Information Processing. Minsk, Belarus (2011)
37. Verma, A., Liu, C.: 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, New Jersey, USA (2011)
38. Verma, A., Liu, C., Jia, J.: New color SIFT descriptors for image classification with applications to biometrics. Int. Journal of Biometrics **1**(3), 56–75 (2011)
39. Verma, A., S., B., Liu, C.: A new color SIFT descriptor and methods for image category classification. In: International Congress on Computer Applications and Computational Science, pp. 819–822. Singapore (2010)
40. Weber, M., Welling, M., Perona, P.: Towards automatic discovery of object categories. In: IEEE Conf. on Computer Vision and Pattern Recognition, vol. 2, pp. 2101–2109. Hilton Head, SC (2000)
41. Yang, J., Liu, C.: Color image discriminant models and algorithms for face recognition. IEEE Trans. on Neural Networks **19**(12), 2088–2098 (2008)
42. Zhang, J., Marszalek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: A comprehensive study. Int. Journal of Computer Vision **73**(2), 213–238 (2007)