Pattern Anal Applic (2015) 18:441–450
DOI 10.1007/s10044-014-0427-1

INDUSTRIAL AND COMMERCIAL APPLICATION

CSIFT based locality-constrained linear coding for image classification

Junzhou Chen · Qing Li · Qiang Peng · Kin Hong Wong

Received: 29 June 2013 / Accepted: 20 October 2014 / Published online: 7 November 2014

Abstract In the past decade, SIFT descriptor has been witnessed as one of the most robust local invariant feature descriptors and widely used in various vision tasks. Most traditional image-classification systems depend on the gray-based SIFT descriptors, which only analyze the gray level variations of the images. Misclassification may happen since their color contents are ignored. In this article, we concentrate on improving the performance of existing image-classification algorithms by adding color information. To achieve this purpose, different kinds of colored SIFT descriptors are introduced and implemented. locality-constrained linear coding (LLC), a state-of-the-art sparse coding technology, is employed to construct the image-classification system for the evaluation. Moreover, we propose a simple ℓ_2-norm regularized local distance to improve the traditional LLC method. The real experiments are carried out on several benchmarks. With the enhancements to color SIFT and ℓ_2-norm regularization, the proposed image-classification system obtains approximately 2 % improvement of classification accuracy on the Caltech-101 dataset and approximately 5 % improvement of classification accuracy on the Caltech-256 dataset.

Keywords CSIFT · Sparse coding · LLC · Image classification

This work is supported by the National Natural Science Foundation of China (No. 61003143) and Sc. & Tech. Plan Project of Sichuan Province China (2012FZ0004).

J. Chen (✉) · Q. Li · Q. Peng
School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, Sichuan, China
e-mail: jzchen@swjtu.edu.cn

Q. Li
e-mail: liqing1988@my.swjtu.edu.cn

Q. Peng
e-mail: qpeng@swjtu.edu.cn

K. H. Wong
Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong
e-mail: khwong@cse.cuhk.edu.hk

1 Introduction

Scale invariant feature transform (SIFT) descriptors [1] are widely used in many vision tasks, such as object recognition, image classification, video retrieval, etc. It has been witnessed a very robust local invariant feature descriptors in respect of different geometrical changes. However, SIFT was mainly developed for gray images; the color information of the objects is neglected. Therefore, two objects with completely different colors may be regarded as the same. To overcome this limitation, different kinds of *Colored SIFT* (CSIFT) descriptors were proposed and developed by researchers to utilize the color information inside the SIFT descriptors [2–6]. With the enhancement of color information, CSIFT descriptors can achieve better performances in resisting certain photometric changes. One example can be found in [3], which shows that CSIFT is more stable than SIFT in case of illumination changes.

On the other hand, the *bag-of-features* (BoF) [7, 8] joined with the *spatial pyramid matching* (SPM) kernel [9] has been employed to build the recent state-of-the-art image-classification systems. In BoF, images are considered as sets of unordered local appearance descriptors, which are clustered into discrete visual words for the representation of images in semantic classification.

SPM divides an image into $2^l \times 2^l$ segments in different scales $l = 0, 1, 2$, computes the BoF histogram within each

segment, and finally concatenates all the histograms to build a spatial location-sensitive descriptor of the image. In order to obtain better classification performance, a codebook (a set of visual words), also named dictionary, is constructed to represent the extracted descriptors. Traditional SPM uses clustering techniques like K-means *vector quantization* (VQ) to generate the codebook. Despite their efficiency, the obtained codebooks usually suffer from several drawbacks such as distortion errors and low discriminative ability [10]. A linear *SPM based on sparse coding* (ScSPM) method [11] was proposed by Yang et al. for relaxing the restrictive cardinality constraint of VQ. By generalizing vector quantization to sparse coding followed by multi-scale spatial max-pooling, ScSPM significantly outperforms the traditional SPM kernel on histograms and is even better than the nonlinear SPM kernels on several benchmarks.

Yu et al. [12] demonstrated that under certain assumptions, locality is more essential than sparsity for the training of nonlinear classifiers and proposed a modification of SC, named *local coordinate coding* (LCC). However, in both SC and LCC, the computationally expensive ℓ_1-norm optimization problem is to be solved. Wang et al. [13] developed a faster implementation of LCC, named *locality-constrained linear coding* (LLC), which utilizes the locality constraint to project each descriptor into its local-coordinate system. It achieves the state-of-the-art image classification accuracy even by just using a linear SVM classifier.

According to our literature survey, although various kinds of *final representation* (fR) based image-classification algorithms with state-of-the-art performances have been developed, most of them use only gray-based SIFT descriptors [10, 11, 13–16]. Using color information can improve the robustness of traditional SIFT descriptor in respect of color variations and the geometrical changes. However, facing the diverse CSIFT descriptors, the following questions are worthwhile to be studied.

- Which CSIFT descriptor is the best for the FR-based image classification system?
- To what extent, the performance of fR-based image classification system can be improved by using CSIFT?

To fully exploit the potential of CSIFT descriptors for image category recognition tasks, a CSIFT-based image-classification system is constructed in this work. As a widely used state-of-the-art SC-based encoding algorithm, LLC is employed to encode the CSIFT descriptors for classification. Moreover, a simple ℓ_2-norm regularized locality distance method is proposed to enhance the performance of traditional LLC.

Real experiments with different kinds of CSIFT descriptors demonstrate that significant improvements can be obtained with the enhancement of color information and ℓ_2-norm regularized locality distance even by only using linear SVM classifier.

The rest of this article is organized as follows: In Sect. 2, a reflectance model for color analysis is presented. In Sect. 3, different kinds of the CSIFT descriptors and their properties are discussed. Section 4 introduces the basic concepts of the LLC. In Sect. 5, we introduce a ℓ_2-norm regularized locality distance method. In Sects. 6 and 7, real experiments are carried out to study the proposed algorithm in various aspects. Finally, in Sect. 8, conclusions are drawn.

2 Dichromatic reflectance model

A physical model of reflection, named *dichromatic reflection model*, was presented by Shafer in 1985 [17] in which the relationship between RGB-values of captured images and the photometric changes, such as shadows and specularities, of environment was investigated. Shafer indicated that the reflection of a incident light can be divided into two distinct components: specular reflection and body reflection. Specular reflection is when a ray of light hits a smooth surface at certain angle. The reflection of that ray will reflect at the same angle as the incident ray. The effect of highlight is caused by the specular reflection. Diffuse reflection is when a ray of light hits the surface which will be reflected back in every direction.

Consider an image of an infinitesimal surface patch of some object. Let the red, green and blue sensors with spectral sensitivities be $f_R(\lambda)$, $f_G(\lambda)$ and $f_B(\lambda)$ respectively. The corresponding sensor values of the surface image are [17, 18]:

$$\begin{aligned} L(\lambda, \mathbf{n}, \mathbf{s}, \mathbf{v}) = m_b(\mathbf{n}, \mathbf{s}) \int_\lambda f_L(\lambda) e(\lambda) c_b(\lambda)\, d\lambda \\ + m_s(\mathbf{n}, \mathbf{s}, \mathbf{v}) \int_\lambda f_L(\lambda) e(\lambda) c_s(\lambda)\, \mathrm{d}\lambda \end{aligned} \tag{1}$$

where $L \in \{R, G, B\}$ is the color channel of light, λ is the wavelength, $\mathbf{n}$ is the surface patch normal, $\mathbf{s}$ is the direction of the illumination source, and $\mathbf{v}$ is the direction of the viewer. $e(\lambda)$ is power of the incident light with wavelength λ, $c_b(\lambda)$ and c_s are the the surface albedo and Fresnel reflectance, respectively. The geometric terms m_b and m_s represent the diffuse reflection and the specular reflection, respectively.

In case white illumination and neutral interface reflection model holds, the incident light energy $e(\lambda) = e$ and Fresnel reflectance term $c_s(\lambda) = c_s$ are both constant values independent of the wavelength λ. By assuming the following holds:

$$\int_\lambda f_R(\lambda) = \int_\lambda f_G(\lambda) = \int_\lambda f_B(\lambda) = f \tag{2}$$

Equation (1) can be simplified:

$$L(\mathbf{n}, \mathbf{s}, \mathbf{v}) = e m_b(\mathbf{n}, \mathbf{s}) k_L + e m_s(\mathbf{n}, \mathbf{s}, \mathbf{v}) c_s f \tag{3}$$

where $k_L = \int_\lambda f_L(\lambda)c_b(\lambda)$ is a variable that depends only on the sensors and the surface albedo.

3 Colored SIFT descriptors

On the basis of the *dichromatic reflection model*, the stability and reliability of color spaces with regard to various photometric events such as shadows and specularities are studied theoretically and empirically [2, 19, 20]. Although there are many existing color space models, they are correlated to intensity; they are linear combinations of RGB; or they are normalized with respect to intensity rgb [19]. In this article, we concentrate on investigating CSIFT using essentially different color spaces: RGB, HSV, YCbCr, Opponent, rg and color invariant spaces.

3.1 SIFT

The SIFT algorithm was originally developed for gray images by Lowe [1, 21] for extracting highly discriminative local image features that are invariant to image scaling and rotation, and partially invariant to changes in illumination and viewpoint. It has been used in a broad range of vision tasks, such as image classification, recognition, content-based image-retrieval, etc. The algorithm involves two steps: (1) extraction of the keypoints of an image and (2) computation of the feature vectors characterizing the keypoints. The first step is carried out by convolving the input image with the DoG (difference of Gaussians) function in multiple scales and detecting the extremas of the outputs. The second step is achieved by sampling the magnitudes and orientations of the image gradient in a patch around the detected feature. A 128-D vector of direction histograms is finally constructed as the descriptor of each patch. Since the SIFT descriptor is normalized, it can invariant to the scale of gradient magnitude. But the light color changes will affect it, because the intensity channel is a combination of the R, G and B channels.

3.2 RGB-SIFT

As the most popular color model, RGB color space provides plenty of information for vision applications. In order to embed RGB color information into the SIFT descriptor, we simply calculate the traditional SIFT descriptors on the each channel of RGB color space. By combining the extracted feature, a 128×3 dimensions descriptor is built (128 for each color channel). Compared with conventional gray-based SIFT, the RGB color gradients (or edges) of the image are captured.

3.3 HSV-SIFT

HSV-SIFT was introduced by Bosch et al. [22] and employed for scene classification task. Similar to RGB SIFT discussed above, they compute SIFT descriptors over all three channels of the HSV color model and produce a 128×3 dimensional SIFT descriptor for each point. It is worth mentioning that H channel of HSV color model is scale-invariant and shift-invariant with respect to light intensity. However, due to the combination of the HSV channels, the entire descriptor has no invariance properties. The conversion from RGB space to HSV space is defined by Eqs. (4)–(6).

$$H = \begin{cases} \text{undefined} & \text{if } \max = \min \\ 60^\circ \times \dfrac{G-B}{\max - \min} + 0^\circ & \text{if } \max = R \text{ and } G \geq B \\ 60^\circ \times \dfrac{G-B}{\max - \min} + 360^\circ & \text{if } \max = R \text{ and } G < B \\ 60^\circ \times \dfrac{G-B}{\max - \min} + 120^\circ & \text{if } \max = G \\ 60^\circ \times \dfrac{G-B}{\max - \min} + 240^\circ & \text{if } \max = B \end{cases} \tag{4}$$

$$S = \begin{cases} 0 & \text{if } \max = 0 \\ \dfrac{\max - \min}{\max} = 1 - \dfrac{\min}{\max} & \text{otherwise} \end{cases} \tag{5}$$

$$V = \max \tag{6}$$

where, max is equal to the maximal one of R, G, B, and min is equal to the minimal one of R, G, B.

3.4 rg-SIFT

The rg-SIFT descriptors are obtained from the rg color space. It is the normalized RGB color model, used r and g channels to describe the color information in the image (b is constant if r and g are given). rg color space is already scale-invariant with respect to light intensity. The conversion from RGB space to rg space is defined as follows,

$$r = \frac{R}{R+G+B} \tag{7}$$

$$g = \frac{G}{R+G+B} \tag{8}$$

3.5 YCbCr-SIFT

As one of the most popular color spaces, YCbCr color space provides very efficient representation of scenes/images and is widely used in the field of video compression. It represents colors in terms of one luminance

component (Y), and two chrominance components (C_b and C_r). The YCbCr-SIFT descriptors are computed on all the channels of YCbCr color space. The YCbCr image can be converted from RGB images using the equation below:

$$\begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.144 \\ -0.1687 & -0.3313 & 0.5 \\ 0.5 & -0.4187 & -0.0813 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} + \begin{bmatrix} 0 \\ 128 \\ 128 \end{bmatrix} \tag{9}$$

3.6 Opponent-SIFT

The opponent color space was first proposed by Ewald Hering in the late nineteenth century [23]. It consists of three channels (O_1, O_2, O_3), in which the O_3 channel represents luminance of the image, while the remainder describe the opponent color (red–green, blue–yellow) of the image. Opponent-SIFT descriptor is obtained by computing the SIFT descriptor over each channel of the opponent color space and combines them together. The RGB images transformed in the opponent color space is defined by Eq. (10).

$$\begin{bmatrix} O_1 \\ O_2 \\ O_3 \end{bmatrix} = \begin{bmatrix} \frac{R-G}{\sqrt{2}} \\ \frac{R+G-2B}{\sqrt{6}} \\ \frac{R+G+B}{\sqrt{3}} \end{bmatrix} \tag{10}$$

3.7 Color invariant SIFT

With the inspiration of the dichromatic reflectance model (see Sect. 2), the color-based photometric invariant scheme was proposed by Geusebroek [2]. It was first applied to SIFT descriptor by Abdel-Hakim and Farag [3]. A linear transformation from RGB to color invariant space is presented as the following:

$$\begin{bmatrix} \hat{E}(x,y) \\ \hat{E}_\lambda(x,y) \\ \hat{E}_{\lambda\lambda}(x,y) \end{bmatrix} = \begin{pmatrix} 0.06 & 0.63 & 0.27 \\ 0.30 & 0.04 & 0.35 \\ 0.34 & 0.60 & 0.17 \end{pmatrix} \begin{bmatrix} R(x,y) \\ G(x,y) \\ B(x,y) \end{bmatrix} \tag{11}$$

where $\hat{E}(x,y)$, $\hat{E}_\lambda(x,y)$, $\hat{E}_{\lambda\lambda}(x,y)$, denote, respectively, the intensity, the yellow–blue channel, and the red–green channel. $\hat{E}$, $\hat{E}_\lambda$ and $\hat{E}_{\lambda\lambda}$ are the spectral differential quotients and represent the same as the above. Measurement of the color invariants is obtained by $\hat{E}$, $\hat{E}_\lambda$ and $\hat{E}_{\lambda\lambda}$.

4 Locality-constrained linear coding

The *bag-of-feature* (BoF) approach has now played a leading role in the field of generic image classification research [11, 13, 15]. It commonly consists of feature extraction, codebook construction, feature coding, and feature pooling. Experimental results shown that, given a visual codebook, choosing an appropriate coding scheme has significant impacts on the classification performance.

Different kinds of coding algorithms are developed [10, 11, 13, 15]; among them, *locality-constrained linear coding* (LLC) [13] is considered as one of the most representative methods, which provides both fast coding speed and state-of-the-art classification accuracy. It has been widely cited in academic papers and employed in image classification applications. In this article, LLC is selected for feature coding in our real experiments.

Let X denote a set of D-dimensional local descriptors in an image, i.e. $X = [x_1, x_2, \ldots, x_N] \in R^{D\times N}$. Let $B = [b_1, b_2, \ldots, b_M] \in R^{D\times M}$ be a visual codebook with M entries. The coding methods convert each descriptor into a M-dimensional code. Unlike the sparse coding, LLC enforces locality constraint instead of sparse constraint. A reconstruction for the basis descriptors B can be acquired by optimizing the following equation:

$$\min_v \sum_{i=1}^{N} \|x_i - Bv_i\|^2 + \lambda \|d_i \odot v_i\|^2 s.t.\ 1^T v_i = 1, \forall i \tag{12}$$

where $\odot$ denotes the element-wise multiplication, and $d_i \in R^M$ is the locality adaptor that gives some degree of freedom for each basis descriptor. LLC ensures these descriptors are proportionally similar to the input descriptor x_i. Specifically,

$$d_i = \exp\left[\frac{\text{dist}(x_i, B)}{\sigma}\right] \tag{13}$$

where $\text{dist}(x_i, B) = [\text{dist}(x_i, b_1), \text{dist}(x_i, b_2), \ldots, \text{dist}(x_i, b_M)]$, and $\text{dist}(x_i, b_j)$ is the Euclidean distance between x_i and b_j. σ is used for adjusting the weight decay speed for the locality adaptor d_i.

An approximation is proposed in [13] to accelerate its computational efficiency in practice by ignoring the second term in Eq. (12). They directly use the K nearest basis descriptors of x_i to minimize the first term. The encoding process is simplified by solving a much smaller linear system,

$$\min_v \sum_{i=1}^{N} \|x_i - Bv_i\|^2 \ s.t.\ 1^T v_i = 1, \forall i \tag{14}$$

This gives the coding coefficients by only selecting k basis vectors. The other coefficients are set to zero.

5 ℓ_2-norm regularized locality distance

In the above section, the details of the traditional LLC method were presented. It can be seen that the distant

function (Eq. 13) plays an important role in the coding scheme. In this paper, we propose a simple ℓ_2 *-norm* regularized locality distance to achieve better classification accuracy. The distance function is defined as:

$$\widetilde{d}_i = \left\| \exp\left[\frac{\mathrm{dist}(x_i, B)}{\sigma} \right] \right\|_{l_2} \quad (15)$$

As a result, Eq. (12) is rewritten as follows:

$$\min_{v} \sum_{i=1}^{N} \|x_i - Bv_i\|^2 + \lambda \|\widetilde{d}_i \odot v_i\|^2 \; s.t. \; 1^T v_i = 1, \forall i \quad (16)$$

6 Experimental results

To evaluate the performances of different kinds of CSIFT descriptors in a *sparse representation* based image classification system, two benchmark datasets: Caltech-101 [24] and Caltech-256 [25] are employed in the real experiment. Since color information is the prerequisite for the CSIFT descriptors computation, to achieve a fair comparison, the gray images in the Caltech-101 and Caltech-256 are removed. To enable that colored images of some categories are sufficient for training a stable classifier (the number of colored images less than 31), we add some new color images of the same category to make sure there are at least 31 colored images in each category.

6.1 Implementation

In all the experiments, the same processing chain with similar the settings is used to ensure consistency.

1. *Colored SIFT* CSIFT/SIFT descriptors extraction. The dense CSIFT/SIFT descriptors are extracted as described in Sect. 3 within a regular spatial grid. The step-size is fixed at 8 pixels and the patch size is fixed at 16×16 pixels. The dimension of gray-based SIFT descriptor is 128. For CSIFT descriptors, RGB-SIFT, SIFT, HSV-SIFT, YCbCr-SIFT, opponent-SIFT, rg-SIFT and color invariance SIFT (C-SIFT) are implemented for the experimentation.
2. Codebooks construction. After the CSIFT/SIFT descriptors are extracted, a codebook of size 1,024 is created using the K-means clustering method on a randomly selected subset (with size 2×10^6) of extracted CSIFT descriptors.
3. *Locality-constrained linear coding* (LLC). The CSIFT/SIFT descriptors are encoded by LLC using the above constructed codebooks. The number of neighbors is set to 5 with the shift-invariant constraint.
4. Pooling with *spatial pyramid matching* (SPM) [9]. The max-pooling operation is adopted to compute the final descriptor of each image. It is performed with a 3 level SPM kernel (1×1, 2×2 and 4×4 sub-regions in the corresponding levels), leaving a same weight at each layer. The pooled features of the sub-regions are concatenated and normalized to form the final descriptor of each image.
5. Classification. A one-versus-all linear SVM classifier [26] is used to train the classifier for its good performances.

6.2 Assessment of color descriptors on the Caltech-101 dataset

The proposed algorithm is carried out using the color images of Caltech-101 dataset, which contains 101 object categories including animals, flowers, vehicles, shapes with significant variance, etc. Some color images are added to avoid insufficient of training data in certain categories as discussed before. The number of original images in every category still varies from 31 to 800. In order to test the performance with different sizes of training data, different numbers (5, 10, . . ., 30) of training images per category are evaluated. In each experiment, we randomly select n images per category for training and leave the remainder for testing. The images were resized to keep the maximum size of height and width no larger than 300 pixels with a conserved aspect ratio. For the sake of simplicity, the codebook size is fixed at 1,024 (the performance of different codebook sizes will be studied in Sect. 7.1). The corresponding results using different kinds of CSIFT descriptors (RGB-SIFT, SIFT, HSV-SIFT, YCbCr-SIFT, opponent-SIFT, rg-SIFT and color invariance SIFT (C-SIFT)) are illustrated in Table 1 and Fig. 1. According to the experimental results, all the CSIFT/SIFT descriptors achieve their best classification accuracy with 30 training images per class. It indicates that more training data may bring better classification accuracy in testing, while the improvement became slight when the size of the number of training images is more than 20. Both RGB-SIFT and YCbCr-SIFT outperform the state-of-the-art gray-based SIFT on this dataset. The YCbCr-SIFT achieves the best performance. For instance, when 30 images of each category are used for training, YCbCr-SIFT obtains the average classification accuracy of 69.1 %; RGB-SIFT provides the second best average classification accuracy (68.6 %). It is worth mentioning that even without color information, SIFT achieves third best average classification accuracy of 68.17 %. Approximately 1 % improvement in average classification accuracy can be obtained by employing CSIFT descriptors.

Table 1 Classification rate (%) comparison on Caltech-101

Training images	5	10	15	20	25	30
RGB-SIFT	45.77 ± 1.02	55.90 ± 0.69	61.26 ± 0.84	64.84 ± 0.68	66.70 ± 0.81	68.65 ± 1.13
SIFT	45.01 ± 0.76	55.39 ± 0.42	60.51 ± 0.60	64.25 ± 0.72	66.29 ± 0.71	68.17 ± 0.98
HSV-SIFT	33.96 ± 0.96	44.06 ± 0.40	50.48 ± 0.60	54.42 ± 0.63	57.76 ± 0.94	59.47 ± 1.31
YCbCr-SIFT	**46.48 ± 0.91**	**56.97 ± 0.60**	**62.09 ± 0.31**	**65.45 ± 0.63**	**68.17 ± 0.76**	**69.18 ± 1.19**
Opponent-SIFT	27.00 ± 0.48	35.07 ± 0.58	39.31 ± 0.55	41.93 ± 0.99	44.21 ± 1.06	45.87 ± 0.74
rg-SIFT	32.51 ± 0.56	41.70 ± 0.88	46.82 ± 0.48	50.35 ± 0.40	53.15 ± 0.83	55.18 ± 1.09
C-SIFT	32.67 ± 0.52	41.90 ± 0.43	47.87 ± 0.56	51.02 ± 0.59	54.05 ± 0.69	55.72 ± 0.88

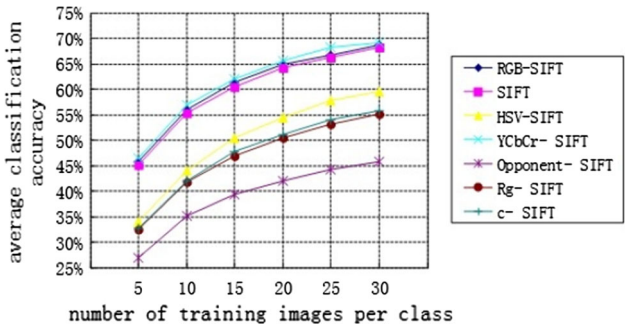

Fig. 1 The different numbers of training images per class on the classification performance

6.3 Assessment of color descriptors on the Caltech-256 dataset

A more complex dataset, Caltech-256 [25], is also employed for the experiments. It consists of 256 object classes and a total of 30,607 images, which have much higher intra-class variability and object location variability compared with the images in Caltech-101. Similar to Sect. 6.2, the gray images are also removed for fair comparison of various CSIFT/SIFT descriptors. Since there are at least 80 color images per category, no more image is added.

In each experiment, we randomly select n ($n \in \{15, 30, 45, 60\}$ is fixed for each experiment) images from every category for training and leave the remainder for testing. For the sake of simplicity, the codebook size is fixed at 4,096 (according to our experience, it produces the best classification performance). The images were resized to keep the maximum size of height and width no larger than 300 pixels with conserved aspect ratio. The details of classification results are shown in Table 2 and Fig. 2. Among all these descriptors, YCbCr-SIFT produces the best performance as well. In case 60 random selected training images of each category are used, YCbCr-SIFT achieves the average classification accuracy of 41.3 %; moreover, RGB-SIFT also provides the second best average classification accuracy (38.7 %). Compared with the performance of gray-based SIFT descriptors, CSIFT brought approximately 4 % enhancement with regard to average classification accuracy, which can be significant in many image classification tasks.

6.4 Assessment of ℓ_2-norm regularized locality distance on the Caltech-101 and Caltech-256 dataset

In Sects. 6.2 and 6.3, different kinds of CSIFT descriptors are implemented and evaluated by traditional LLC method. In this section, ℓ_2-norm regularized locality distance and CSIFT descriptors are combined together to obtain better performance. The datasets we used are the same as in Sects. 6.2 and 6.3. Since YCbCr-SIFT descriptor and RGB-SIFT descriptor achieved the top two classification accuracies, they are employed for comparison. The size of the codebook is set at 1,024. We randomly selected the training images and repeated the experiments 10 times. The corresponding average results are listed in Tables 3 and 4. The YCbCr-SIFT descriptor still provides the best performances. With the enhancement of ℓ_2-norm regularized locality distance, the classification accuracy increases

Table 2 Classification rate (%) comparison on Caltech-256

Training images	15	30	45	60
RGB-SIFT	26.70 ± 0.33	33.04 ± 0.22	36.56 ± 0.32	38.71 ± 0.38
SIFT	25.06 ± 0.07	31.22 ± 0.24	34.92 ± 0.39	37.22 ± 0.35
HSV-SIFT	21.95 ± 0.30	28.18 ± 0.22	31.79 ± 0.28	34.03 ± 0.29
YCbCr-SIFT	**28.58 ± 0.32**	**35.20 ± 0.18**	**38.97 ± 0.34**	**41.31 ± 0.27**
Opponent-SIFT	14.37 ± 0.24	17.92 ± 0.22	20.0 ± 0.20	21.43 ± 0.45
rg-SIFT	18.16 ± 0.24	22.98 ± 0.26	25.88 ± 0.36	27.63 ± 0.31
C-SIFT	14.56 ± 0.18	19.30 ± 0.22	22.13 ± 0.19	24.19 ± 0.27

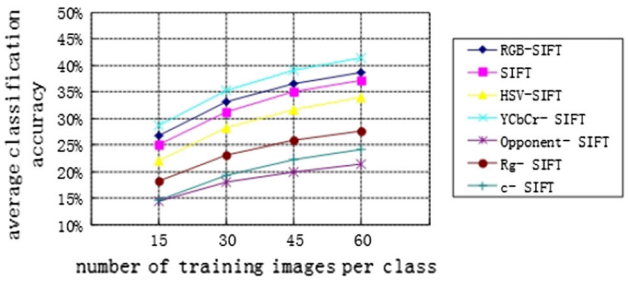

Fig. 2 The different number of training images per class on the classification performance

steadily (about 0.6 %). In Table 3, when 30 training images of each category are used, YCbCr-SIFT achieves the average classification accuracy of 69.74 %; approximately, 1.57 % enhancement is obtained. In Table 4, when 60 training images of each category are used, YCbCr-SIFT achieves the average classification accuracy of 41.78 %; the combination method obtains approximately 4.56 % enhancement compared with traditional LLC.

7 Further evaluations

The experimental results in Sects. 6.2 and 6.3 show that, among the different CSIFT descriptors, YCbCr-SIFT and RGB-SIFT achieve better image classification performance than the state-of-the-art gray-based SIFT. While, it is well known that choosing different codebook sizes, different numbers of neighbors in LLC and different pooling methods will affect the final classification results. In this section, further evaluations are carried out for more comprehensive studies of these two CSIFT descriptors.

7.1 Impact of codebook size

Firstly, we test the impacts of different codebook sizes (512, 1,024 and 2,048) using the Caltech-101 dataset. As discussed in Sect. 6, the codebooks are trained by the *K*-means clustering algorithm. Different numbers (5, 10, ..., 30) of training images per category are evaluated.

Table 3 Classification rate (%) comparison on Caltech-101

Training images	5	10	15	20	25	30
LLC						
RGB-SIFT	45.77 ± 1.02	55.90 ± 0.69	61.26 ± 0.84	64.84 ± 0.68	66.70 ± 0.81	68.65 ± 1.13
SIFT	45.01 ± 0.76	55.39 ± 0.42	60.51 ± 0.60	64.25 ± 0.72	66.29 ± 0.71	68.17 ± 0.98
YCbCr-SIFT	**46.48 ± 0.91**	**56.97 ± 0.60**	**62.09 ± 0.31**	**65.45 ± 0.63**	**68.17 ± 0.76**	**69.18 ± 1.19**
LLC + ℓ_2-norm						
RGB-SIFT	45.94 ± 0.84	55.92 ± 0.72	61.33 ± 0.83	64.92 ± 0.61	66.77 ± 0.83	68.76 ± 1.14
SIFT	45.01 ± 0.80	55.81 ± 0.40	61.22 ± 0.30	64.09 ± 0.96	66.34 ± 0.88	68.65 ± 1.1
YCbCr-SIFT	**47.18 ± 0.91**	**57.39 ± 0.46**	**62.41 ± 0.56**	**65.98 ± 0.51**	**68.17 ± 0.68**	**69.74 ± 0.87**

Table 4 Classification rate (%) comparison on Caltech-256

Training images	15	30	45	60
LLC				
RGB-SIFT	26.70 ± 0.33	33.04 ± 0.22	36.56 ± 0.32	38.71 ± 0.38
SIFT	25.06 ± 0.07	31.22 ± 0.24	34.92 ± 0.39	37.22 ± 0.35
YCbCr-SIFT	**28.58 ± 0.32**	**35.20 ± 0.18**	**38.97 ± 0.34**	**41.31 ± 0.27**
LLC + l_2-norm				
RGB-SIFT	27.26 ± 0.15	33.32 ± 0.18	36.91 ± 0.19	39.06 ± 0.39
SIFT	26.54 ± 0.35	32.59 ± 0.13	35.85 ± 0.33	38.20 ± 0.39
YCbCr-SIFT	**29.24 ± 0.28**	**35.68 ± 0.25**	**39.29 ± 0.18**	**41.78 ± 0.44**

Table 5 The codebooks of size 512

Training images	5	10	15	20	25	30
SIFT	46.01 ± 0.65	55.81 ± 0.41	60.98 ± 0.50	63.99 ± 0.97	66.23 ± 0.49	67.10 ± 1.10
RGB-SIFT	46.57 ± 0.59	56.28 ± 0.60	60.92 ± 0.45	64.10 ± 0.62	66.01 ± 0.82	67.10 ± 1.26
YCbCr-SIFT	**46.81 ± 0.81**	**57.18 ± 0.39**	**62.25 ± 0.56**	**65.53 ± 0.65**	**67.62 ± 0.61**	**69.16 ± 0.80**

Table 6 The codebooks of size 1,024

Training images	5	10	15	20	25	30
SIFT	45.01 ± 0.76	55.39 ± 0.42	60.51 ± 0.60	64.25 ± 0.72	66.29 ± 0.71	68.17 ± 0.98
RGB-SIFT	45.77 ± 1.02	55.90 ± 0.69	61.26 ± 0.84	64.84 ± 0.68	66.70 ± 0.81	68.65 ± 1.13
YCbCr-SIFT	**46.48 ± 0.91**	**56.97 ± 0.60**	**62.09 ± 0.31**	**65.45 ± 0.63**	**68.17 ± 0.76**	**69.18 ± 1.19**

Table 7 The codebooks of size 2,048

Training images	5	10	15	20	25	30
SIFT	43.56 ± 0.78	54.18 ± 0.78	60.08 ± 0.72	63.18 ± 0.54	65.68 ± 0.63	67.91 ± 1.21
RGB-SIFT	43.79 ± 0.91	54.33 ± 0.55	59.89 ± 0.73	63.07 ± 0.94	65.77 ± 0.73	67.94 ± 0.79
YCbCr-SIFT	**44.62 ± 0.75**	**55.21 ± 0.51**	**61.42 ± 0.33**	**65.13 ± 0.66**	**67.42 ± 0.64**	**69.45 ± 0.84**

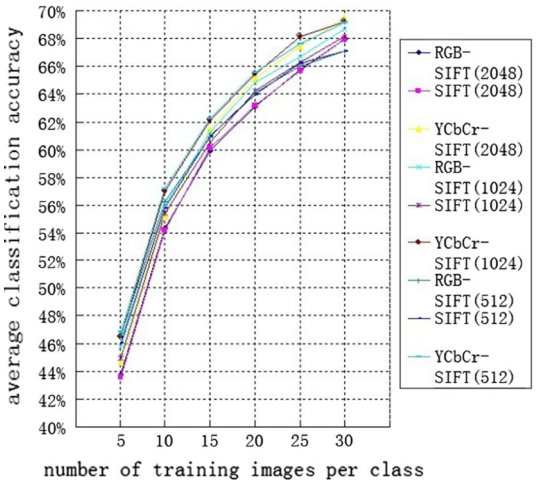

Fig. 3 The different numbers of training images per class on the classification performance

The number of neighbors in LLC is set at 5. The corresponding results are presented in Tables 5, 6, 7 and Fig. 3. YCbCr-SIFT descriptor outperforms the others in all the tests. In most cases, the highest classification accuracy is obtained by using codebook of size 1,024. However, when the codebook of size 2,048 is utilized, the classification accuracies decrease (except YCbCr-SIFT descriptor with 30 training images per category). It may be caused by the over-completeness of the codebooks, which results in large deviations in representing similar local features. It is interesting to notice that, by using more training data, the problem of over-completeness might be overcome. For instance, YCbCr-SIFT descriptor with codebooks of size 2,048 and 30 training images per category achieves the highest average classification accuracy.

7.2 Impact of different numbers of neighbors

The performances of the proposed algorithm using different numbers of neighbors K in LLC are also estimated. The codebook size is fixed at 1,024, and the number of training images per category is 30. The results are shown in Table 8 and Fig. 4. With the increase of the neighbor number K in LLC, the classification accuracy takes on the trend of rising first, then drops after $K \geq 25$. The highest average classification accuracy is obtained by using YCbCr-SIFT descriptor (72.59 %). In contrast to the highest classification result of SIFT (69.18 %), more than 3 % improvement is achieved.

7.3 Comparison of pooling methods

Besides the max-pooling method, sum-pooling is another choice which can also be used to summarize the features of each SPM layer. Tables 9 and 10 show the experimental results using the two methods, respectively. In Fig. 5 they are illustrated together for comparison. The codebook size is 1,024. The number of neighbors used in LLC is 5. It can be noticed that the max-pooling method significantly outperforms sum-pooling.

Table 8 Comparison on the sizes of the neighborhood size

Number of K	5	10	15	20	25	30
SIFT	67.91 ± 1.21	68.41 ± 1.03	68.74 ± 0.94	68.31 ± 0.84	68.99 ± 0.86	68.51 ± 1.17
RGB-SIFT	67.94 ± 0.79	68.61 ± 0.82	68.72 ± 0.89	68.99 ± 0.71	69.18 ± 1.1	68.78 ± 0.13
YCbCr-SIFT	**69.45 ± 0.84**	**70.44 ± 1.03**	**71.37 ± 0.72**	**72.59 ± 0.63**	**72.56 ± 1.22**	**72.39 ± 1.47**

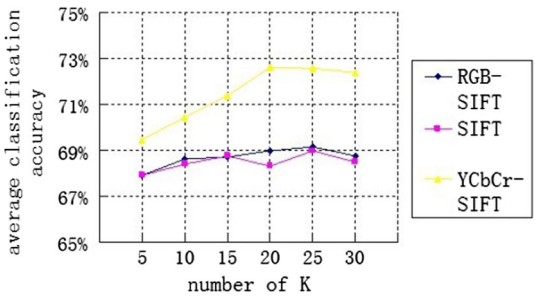

Fig. 4 The different numbers of training images per class on the classification performance

$$\text{Max} : v_j = \max(v_1, v_2, \ldots, v_i) \tag{17}$$

$$\text{Sum} : v_j = v_1 + v_2 + \cdots + v_i \tag{18}$$

As can be seen from Fig. 5, the best performance is achieved by the combination of "*max-pooling*" and "ℓ_2-*normalization*".

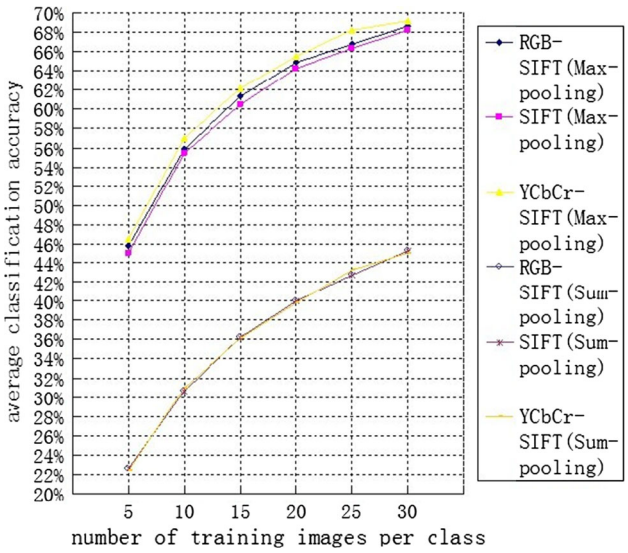

Fig. 5 Impact of different pooling methods

8 Conclusion

In this article, CSIFT descriptors are introduced to improve the state-of-the-art *locality-constrained linear coding* (LLC) based image classification system. Different kinds of CSIFT descriptors are implemented and evaluated with varying settings of the parameters. Real experiments have demonstrated that, by utilizing color information, considerable improvements can be obtained. Among the CSIFT descriptors, YCbCr-SIFT descriptor achieves the most stable and accurate image classification performance. Compared with the highest average classification accuracy achieved by using gray-based SIFT descriptors, YCbCr-SIFT descriptor acquired approximately 1 % increase on the Caltech-101 dataset (see Sect. 7.2) and approximately 4 % increase on the Caltech-256 dataset (see Sect. 6.3). Besides the YCbCr-SIFT descriptor, RGB-SIFT descriptor also provides favorable performance. As one of the most representative FR-based image-classification algorithms, the improvements achieved on LLC show that using CSIFT descriptors is an approach with good potential to enhance state-of-the-art FR-based image-classification systems. On the other hand, although be reported can achieve invariant or discriminatory object recognition, we found that the performances of some others CSIFT descriptors are not as good as expected. On the other hand, although be reported can achieve invariant or discriminatory object recognition, we found that the performances of some others CSIFT descriptors are not as good as expected. Moreover, we obtain a steady rise in the classification accuracy by introducing a simple ℓ_2 *-norm* regularized locality distance.

Table 9 The performance of max-pooling

Training images	5	10	15	20	25	30
SIFT	45.01 ± 0.76	55.39 ± 0.42	60.51 ± 0.60	64.25 ± 0.72	66.29 ± 0.71	68.17 ± 0.98
RGB-SIFT	45.77 ± 1.02	55.90 ± 0.69	61.26 ± 0.84	64.84 ± 0.68	66.70 ± 0.81	68.65 ± 1.13
YCbCr-SIFT	**46.48 ± 0.91**	**56.97 ± 0.60**	**62.09 ± 0.31**	**65.45 ± 0.63**	**68.17 ± 0.76**	**69.18 ± 1.19**

Table 10 The performance of sum-pooling

Training images	5	10	15	20	25	30
SIFT	22.14 ± 0.78	30.14 ± 0.85	**36.38 ± 0.47**	38.98 ± 1.03	41.86 ± 0.61	45.0 ± 1.06
RGB-SIFT	**22.67 ± 0.73**	30.64 ± 0.63	36.26 ± 0.87	**40.04 ± 0.41**	42.71 ± 0.82	**45.24 ± 0.77**
YCbCr-SIFT	22.42 ± 1.06	**31.04 ± 0.65**	36.12 ± 0.62	39.83 ± 0.83	**43.28 ± 0.87**	45.10 ± 1.33

The combination of YCbCr-SIFT descriptor and the ℓ_2 -*norm* regularized locality distance provides the best performance. It achieves approximately 2 % improvement of classification accuracy on the Caltech-101 dataset and approximately 5% improvement of classification accuracy on the Caltech-256 dataset. Our future work will investigate the combinations of learning-based color descriptors [27], different kinds of distance functions and sparse coding technologies to achieve better image classification performance.

Acknowledgments The authors would like to thank the anonymous reviewers for their comments and suggestions to improve this article.

References

1. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
2. Geusebroek J-M, van den Boomgaard R, Smeulders AWM, Geerts H (2001) Color invariance. IEEE Trans Pattern Anal Mach Intell 23(12):1338–1350
3. Abdel-Hakim AE, Farag AA (2006) Csift: a sift descriptor with color invariant characteristics. In: 2006 IEEE computer society conference on computer vision and pattern recognition, vol 2. IEEE, pp 1978–1983
4. Van De Weijer J, Gevers T, Bagdanov AD (2006) Boosting color saliency in image feature detection. IEEE Trans Pattern Anal Mach Intell 28(1):150–156
5. Burghouts GJ, Geusebroek J-M (2009) Performance evaluation of local colour invariants. Comput Vis Image Underst 113(1):48–62
6. Gevers T, Gijsenij A, Van de Weijer J, Geusebroek J-M (2012) Color in computer vision: fundamentals and applications, vol 24. Wiley, New York
7. Goldfarb D, Idnani A (1983) A numerically stable dual method for solving strictly convex quadratic programs. Math Program 27(1):1–33
8. Csurka G, Dance C, Fan L, Willamowski J, Bray C (2004) Visual categorization with bags of keypoints. In: Workshop on statistical learning in computer vision, ECCV, vol 1, pp 22
9. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: 2006 IEEE computer society conference on computer vision and pattern recognition, vol 2. IEEE, pp 2169–2178
10. Shabou A, LeBorgne H (2012) Locality-constrained and spatially regularized coding for scene categorization. In: 2012 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 3618–3625
11. Yang J, Yu K, Gong Y, Huang T (2009) Linear spatial pyramid matching using sparse coding for image classification. In: IEEE conference on computer vision and pattern recognition, 2009. CVPR 2009. IEEE, pp 1794–1801
12. Yu K, Zhang T, Gong Y (2009) Nonlinear learning using local coordinate coding. Adv Neural Inf Process Syst 22:2223–2231
13. Wang J, Yang J, Yu K, Lv F, Huang T, Gong Y (2010) Locality-constrained linear coding for image classification. In: 2010 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 3360–3367
14. Yang J, Yu K, Huang T (2010) Supervised translation-invariant sparse coding. In: 2010 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 3517–3524
15. Liu L, Wang L, Liu X (2011) In defense of soft-assignment coding. In: 2011 IEEE international conference on computer vision (ICCV). IEEE, pp 2486–2493
16. Yang M, Zhang L, Feng X, Zhang D (2011) Fisher discrimination dictionary learning for sparse representation. In: 2011 IEEE international conference on computer vision (ICCV). IEEE, pp 543–550
17. Shafer SA (1985) Using color to separate reflection components. Color Res Appl 10(4):210–218
18. Gevers T, Van De Weijer J, Stokman H et al (2007) Color feature detection. Color image processing: methods and applications, pp 203–226
19. Gevers T, Smeulders WM et al (1999) Color based object recognition. Pattern Recognit 32(3):453–464
20. van de Sande KEA, Gevers T, Snoek CGM (2010) Evaluating color descriptors for object and scene recognition. IEEE Trans Pattern Anal Mach Intell 32(9):1582–1596
21. Lowe DG (1999) Object recognition from local scale-invariant features. In: The proceedings of the seventh IEEE international conference on computer vision, vol 2. IEEE, pp 1150–1157
22. Bosch A, Zisserman A, Muoz X (2008) Scene classification using a hybrid generative/discriminative approach. IEEE Trans Pattern Anal Mach Intell 30(4):712–727
23. Hering E (1964) Outlines of a theory of the light sense, vol 344. Harvard University Press, Cambridge
24. Fei-Fei L, Fergus R, Perona P (2006) One-shot learning of object categories. IEEE Trans Pattern Anal Mach Intell 28(4):594–611
25. Griffin G, Holub A, Perona P (2007) Caltech-256 object category dataset
26. Fan R-E, Chang K-W, Hsieh C-J, Wang X-R, Lin C-J (2008) Liblinear: a library for large linear classification. J Mach Learn Res 9:1871–1874
27. Khan R, Van de Weijer J, Khan FS, Muselet D, Ducottet C, Barat C (2013) Discriminative color descriptors. In: 2013 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 2866–2873