Improving Image Classification Accuracy with ELM and CSIFT

Qing Li, Qiang Peng, Junzhou Chen, Chuan Yan

***Abstract*—We present a new image classification system by integrating color information and the extreme learning machine algorithm based on the localized soft-assignment coding classification framework. Our system is original with its unique structure. It improves the image classification accuracy significantly compared to the existing image classification systems. Specifically, on classification accuracy, our system can improve more than 4% on the Caltech-101 dataset and achieve up to 3% improvement on the Caltech- 256 dataset compared to existing systems.**

***Index Terms*—color information, extreme learning machine, image classification.**

I. Introduction

Image classification is a very challenging problem which is to classify images to the specified one or more categories. It has gained more and more attention in computer vision and machine learning. In recent years, the *bag-of-features* (BoF) is the most popular and effective image classification method [1][2]. BoF typically involves obtaining the set of bags of features then creating the histogram using the bags as the bins. This histogram is then used for image classification. BoF neglects the information of the features' spatial distribution, which severely limits the description ability of the image representation. Hence, to overcome this shortcoming, Lazebnik *et al*. [3] proposed a particular method, namely the *spatial pyramid matching* (SPM) method, which can be integrated with BoF. SPM can separate an image into more spatial sub-regions [4]. For instance, it can divide an image into $2^l \times 2^l$ segments in different scales, where $l = 0, 1, 2$. After that, it computes the BoF histogram from each segment, and finally concatenates all the histograms to build a spatial location sensitive vector of the image.

Qing Li was in the School of Information Science & Technology, Southwest Jiaotong University, Chengdu, Sichuan, 610031, P.R. China email: (liqing1988@my.swjtu.edu.cn).

Qiang Peng was in the School of Information Science & Technology, Southwest Jiaotong University, Chengdu, Sichuan, 610031, P.R. China email: (pqiang@nec.swjtu.edu.cn).

Junzhou Chen was in the School of Information Science & Technology, Southwest Jiaotong University, Chengdu, Sichuan, 610031, P.R. China email: (jzchen@swjtu.edu.cn).

Chuan Yan was in the School of Information Science & Technology, Southwest Jiaotong University, Chengdu, Sichuan, 610031, P.R. China email: (kirin@my.swjtu.edu).

The framework of SPM based on BoF has been applied to build image classification systems recently [2], which is shown in Figure 1 (a). It contains five steps: 1) feature extraction - extract patches, and obtain the descriptors, and apply statistical analysis on the feature points. 2) codebook creation - construct a codebook, a set of visual words, from the extracted descriptors. The codebook is also named the dictionary; 3) feature coding - represent each feature descriptor by a codeword. The coding process adopts the *vector quantization* (VQ) method; 4) spatial pooling - pool the codes of each sub-region together in a SPM layer, and then concatenate all sub-regions to the global image representation; and 5) classification - send the final representation to a classifier to obtain the result.

In order to obtain better classification performance, researchers aim at studying the feature coding method. Hard voting is the original coding method, which reflects the code's occurrence frequency. Despite it is fast and simple, ambiguous information is often neglected and large quantization error is produced [5]. To relieve this issue, soft-assignment is proposed [6]. It estimates the membership of a local feature to different visual words [5][1]. A linear *SPM method based on sparse coding* (ScSPM) was proposed by Yang et al. [4] to relax the restrictive cardinality constraint of VQ. ScSPM significantly outperforms the traditional SPM on histograms and is even better than the nonlinear SPM. Yu *et al*. [7] observed that the results of *sparse coding* (SC) tend to assign nonzero coefficients to bases in the encoded data. They proposed a modification to SC, named *Local Coordinate Coding* (LCC). However, similar to SC, LCC takes too much time to solve the L1-norm optimization problem. To address this problem, Wang *et al*. [8] developed a faster LCC implementation, namely *locality-constrained linear coding* (LLC). It utilizes the locality constraint to project each descriptor into its local-coordinate system, and the projected coordinates are integrated with max pooling to generate the final representation. Based on the traditional soft-coding method,

Digital Object Identifier 10.1109/MCSE.2018.108164708 1521-9615/$26.00 2018 IEEE

Liu *et al*. [1] roposed a *localized soft-assignment coding* (LSC). This coding method keeps up with or even outperforms the sparse and local coding schemes. Although feature coding method have been extensively conducted to enhance classification performance, few works have been explored for feature descriptors or classifier, which is the focus of this work.

It is acknowledged that the SIFT, as one of the most successful feature descriptors, is efficiently used in the BoF framework. *Scale-invariant feature transform* (or *SIFT*) is an algorithm to identify local features in images. For any object in an image, interesting points on the object can be extracted to provide a "feature description" of the object, which can then be used to identify the same object in another image. To perform reliable recognition, the features extracted from the training image are detectable in the test images with different scale, noise, and illumination. However, SIFT was only used for gray images, and color information in color images are ignored. If the prominent classification information are the colors, SIFT may not provide enough distinguishable ability in image classification tasks, so substantial misclassification cases may occur. For this reason, different colored scale invariant feature transform (CSIFT) algorithms, which utilize the color information in addition to SIFT, were proposed [9]. In this paper, we focus primarily on improving the performance of the traditional BoF framework by integrating color information into the classification system. To accomplish this task, various kinds of colored SIFT descriptors were proposed and implemented in the BoF framework.

Generally, linear SVM is adopted for the final classification. However, we use a novel method, namely the *extreme learning machine* (ELM) [10], to build a challenging image classification framework. Although ELM was originally proposed for the *single hidden-layer feed forward neural networks* (SLFNs). ELM can effectively avoid the slow training speed and overfitting problems suffered by traditional neural network training algorithms [11]. Meanwhile, ELM has obtained the higher generalization performance at a much faster speed in many applications, which are used for both the regression and the classification problem. Therefore, we use the ELM method, in addition to the traditional BoF framework, to enhance the image classification performance.

In this paper, we present a new image classification system in order to improve the classification performance. First, we construct CSIFT descriptors-based image classification system for image category tasks, so that we can exploit the color information. Among the existing coding schemes, *localized soft-assignment coding* (LSC) is a widely used state-of-the-art encoding algorithm, which is employed to encode our CSIFT descriptors. As others, we adopt the linear SVM classifier to achieve better performances. Second, we also bring in a new ELM learning algorithm to compare our framework to the traditional BoF framework. Our framework can obtain better classification accuracy. In the end, we achieve better performance with the new framework, boosting LSC with CSIFT and ELM classifiers. Our image classification framework is shown in Figure 1 (b), which is original with its unique structure. It improves the image classification accuracy significantly compared to the existing image classification systems. Specifically, our system can improve more than 4% on classification accuracy on the Caltech-101 dataset, and achieve up to 3% improvement on the Caltech-256 dataset compared to existing systems.

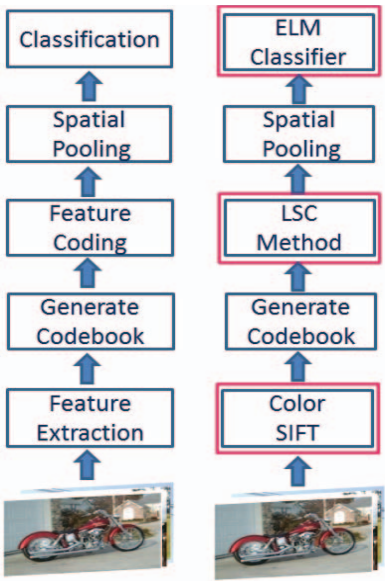

Fig. 1. (a) The current advanced image classification system framework. (b) Our new image classification framework.

II. Localized Soft-assignment Coding

Recent on studies show that, as one core step, feature coding influences the image classification performance. In a recent review, various coding algorithms have been proposed that lead to better classification and recognition performance [12]. Among these coding schemes, LSC is one of the most representative methods that provides the state-of-the-art classification performance [13]. It is based on a traditional soft-coding method by adding

Digital Object Identifier 10.1109/MCSE.2018.108164708 1521-9615/$26.00 2018 IEEE

the locality constraint in the distance function. In our implementation, LSC is applied for feature coding.

Let X denotes a set of D-dimensional local descriptors in an image, i.e. $X = [x_1, x_2, \ldots, x_N] \in R^{D\times N}$. Given a visual codebook with M entries, $B = [b_1, b_2, \ldots, b_M] \in R^{D\times M}$. The coding methods convert each descriptor into an M-dimensional code to generate the final image representation. Because the local features in generic image classification are often similar rather than identical, it will produce high sensitivity to adversely affect the similarity estimate and in turn the coding result. To solve this problem, the LSC scheme employs the *k* visual words in the neighborhood of a local feature and conceptually set its distances to the remaining words as infinity. Formally, x_i denotes a local feature and b_j denotes the *j*th visual word. Here, *k* is our coding coefficient. Then, the localized soft-assignment coding is:

$$v_{ij} = \frac{exp(-\beta \hat{D}(x_i, b_j))}{\sum_{n=1}^{M} exp(-\beta \hat{D}(x_i, b_n))},$$
$$\hat{D}(x_i, b_n) = \begin{cases} dist(x_i, b_n) & \text{if } b_n \in N_k(x_i) \\ \propto & otherwise \end{cases} \tag{1}$$

where β is a smoothing factor controlling how widely the assignment distributes the weights across all the codewords, $\hat{D}$ denotes the localized version of the original distance $dist(x_i, b_n)$, $dist(x_i, b_n)$ is the Euclidean distance between x_i and b_n, and N_k is the the K nearest basis descriptors of x_i defined by the distance $dist(x_i, b_n)$, and $n = 1, 2, ..., M$.

III. Extreme learning machines

Extreme learning machine has been known as a novel learning paradigm, and it is originally proposed for the single hidden-layer feedforward neural networks (SLFNs). Then the learning method is extended to the "generalized" SLFNs where the hidden layer does not need to be neuron alike [14].

The essence of ELM is that, different from the common training of a feedforward neural networks, the hidden layer of a SLFNs does not need to be tuned. Given a set of training data (x_i,y_i), $i = 1, ..., N$, where $x_i \in R^d$ and $y_i \in -1, 1$, the output function of the generalized SLFNs with L hidden nodes, the ELM model can be written as follows:

$$f_L(x) = \sum_{i=1}^{L} \alpha_i h_i(x) = \mathbf{h}(\mathbf{x})\boldsymbol{\alpha} \tag{2}$$

where $\boldsymbol{\alpha} = [\alpha_1, ..., \alpha_L]^T$ is the connection weights between the hidden layer of L nodes and the output node; $\mathbf{h}(\mathbf{x}) = [h_1(x), ..., h_L(x)]$ is the hidden layer output matrix of the SLFN, and the row of $\mathbf{h}(\mathbf{x})$ is the hidden neuron output with respect to the input $\mathbf{x}$. $\mathbf{h}(\mathbf{x})$ actually maps the data from the d-dimensional input space to the L-dimensional hidden-layer feature space (ELM feature space) H, and thus, $\mathbf{h}(\mathbf{x})$ is indeed a feature mapping.

Compared to the traditional learning algorithms [14], ELM not only tends to attain the lowest training error, but also the lowest norm of output weights. In accordance with Bartlett's theory, when the feedforward neural networks achieve the least training error, the norms of weights are least, and the universalization performance of the networks is better. It can be seen that ELM is to minimize the training error as well as the norm of the output weights [11],

$$Minimize : \|\mathbf{H}\boldsymbol{\alpha} - \mathbf{T}\|^2 \ and \ \|\boldsymbol{\alpha}\| \tag{3}$$

where $\mathbf{H}$ is the hidden-layer output matrix.

$$\mathbf{H} = \begin{bmatrix} \mathbf{h}(\mathbf{x_1}) \\ \vdots \\ \mathbf{h}(\mathbf{x_N}) \end{bmatrix} = \begin{bmatrix} h_1(x_1) & \ldots & h_L(x_1) \\ \vdots & \ldots & \vdots \\ h_1(x_N) & \ldots & h_L(x_N) \end{bmatrix} \tag{4}$$

Instead of the standard optimization method, the smallest norm least square method was used in the original ELM implementation,

$$\boldsymbol{\alpha} = \mathbf{H}^{\dagger}\mathbf{T} \tag{5}$$

where $\mathbf{H}^{\dagger}$ is the Moore-Penrose generalized inverse of matrix $\mathbf{H}$ [10].

A. Random hidden layer feature mapping based ELM

If ELM makes use of the orthogonal projection method, it can efficiently have two cases: when $\mathbf{H}^T\mathbf{H}$ is nonsingular and $\mathbf{H}^{\dagger} = (\mathbf{H}^T)^{-1}\mathbf{H}^T$, or when $\mathbf{H}\mathbf{H}^T$ is nonsingular and $\mathbf{H}^{\dagger} = \mathbf{H}^T(\mathbf{H}^T)^{-1}$. Based on the ridge regression theory, it was mentioned that a positive value can be added to the diagonal of $\mathbf{H}^T\mathbf{H}$ or $\mathbf{H}\mathbf{H}^T$ in the computing the output weights $\boldsymbol{\alpha}$ [15].

In order to improve the stability of ELM we can have

$$\boldsymbol{\alpha} = \mathbf{H}^T\left(\frac{\mathbf{I}}{\mathbf{C}} + \mathbf{H}\mathbf{H}^T\right)^{-1}\mathbf{T}$$
$$or \quad \boldsymbol{\alpha} = \left(\frac{\mathbf{I}}{\mathbf{C}} + \mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{T} \tag{6}$$

where C is a positive constant. According to the above solution, it has been shown the better universalization

Digital Object Identifier 10.1109/MCSE.2018.108164708 1521-9615/$26.00 2018 IEEE

performance. And the corresponding ELM output function is:

$$f(x) = \mathbf{h}(\mathbf{x})\boldsymbol{\alpha} = \mathbf{h}(\mathbf{x})\mathbf{H}^T\left(\frac{\mathbf{I}}{\mathbf{C}} + \mathbf{H}\mathbf{H}^T\right)^{-1}\mathbf{T}$$
$$or \quad f(x) = \mathbf{h}(\mathbf{x})\boldsymbol{\alpha} = \mathbf{h}(\mathbf{x})\left(\frac{\mathbf{I}}{\mathbf{C}} + \mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{T} \tag{7}$$

B. Kernel based ELM

We should be paying more attention to how nonlinear activation and kernel functions are used in ELM. When the hidden layer feature mapping $\mathbf{h}(\mathbf{x})$ is unknown to users, one can apply Mercer's conditions to define a kernel matrix for ELM as follows:

$$\Omega_{ELM} = \mathbf{H}\mathbf{H}^T : \Omega_{ELM_{ij}} = \mathbf{h}(\mathbf{x_i})\mathbf{h}(\mathbf{x_j}) = \mathbf{K}(\mathbf{x_i}, \mathbf{x_j}) \tag{8}$$

Then we can write the output function of ELM classifier Eq.(7) compactly as:

$$\begin{aligned} f(x) &= \mathbf{h}(\mathbf{x})\boldsymbol{\alpha} \\ &= \mathbf{h}(\mathbf{x})\mathbf{H}^T\left(\frac{\mathbf{I}}{\mathbf{C}} + \mathbf{H}\mathbf{H}^T\right)^{-1}\mathbf{T} \\ &= \begin{bmatrix}\mathbf{K}(\mathbf{x}, \mathbf{x_1}) \\ \vdots \\ \mathbf{K}(\mathbf{x}, \mathbf{x_N})\end{bmatrix}^T\left(\frac{\mathbf{I}}{\mathbf{C}} + \Omega_{ELM}\right)^{-1}\mathbf{T} \end{aligned} \tag{9}$$

The kernel matrixes can be implemented on ELM. That is the hidden layer feature mapping $\mathbf{h}(\mathbf{x})$ need not be known to users, instead its corresponding kernel $\mathbf{K}(\mathbf{u}, \mathbf{v})$ (e.g. $\mathbf{K}(\mathbf{u}, \mathbf{v}) = exp(-\beta \|\mathbf{u} - \mathbf{v}\|)$) is provided to users. As a matter of fact, the number of hidden nodes L (the dimensionality of the hidden layer feature space) need not be specified either.

IV. Experimental Results

In this section, we try to verify that: 1) CSIFT descriptors used in LSC can improve the classification performance; 2) we bring in an ELM learning method instead of the common learning algorithms, which can obtain comparable or even better classification capability; 3) we propose the new final representation framework (CSIFT descriptors and ELM learning are combined with LSC), which can significantly enhance the image classification results.

In the first verification experiment, we used two kinds of the CSIFT descriptors with the LSC method to test the image classification performances, and we report results based on two sets of benchmark datasets: Caltech-101 and Caltech-256.

In paper [9], which is part of our experiments. It has been witnessed that the YCbCr-SIFT is the most stable and accurate image classification performance, and the second is the RGB-SIFT descriptor. The original Caltech-101 dataset includes 9144 images with 101 object categories. Each object category contains about 40 to 800 images. The original Caltech-256 dataset has 256 object categories containing a total of 30607 images. Compared with the Caltesch-101, the minimum number of images in any category is increased to 80 images. Since color information is the prerequisite for CSIFT descriptors computation, to achieve a fair comparison, gray images in the original Caltech-101 and Caltech-256 datasets are removed. In order to make sure there are enough color images in each category for the experiment, we add some more images to each category to constitute colored datasets (the minimum number of images in each category is 40 color images in the colored Caltech-101 dataset and 80 color images in the colored Caltech-256 dataset).

In the second verification experiment, we introduce an ELM learning method, which was combined with LSC to compare with some popular feature coding approaches on the original Caltech-101 dataset.

In the final verification experiment, we evaluate the new alliance of the framework on the colored Caltech-101 dataset.

A. Implementation

All of our experiments take the same processing chain, which are similar to the settings in the referenced papers. Those are used to ensure the consistency of experiments.

1) CSIFT or SIFT descriptors extraction. We use a dense spatial grid to the feature pathes. The step-size is fixed to 6 pixels on Caltech-101 and 8 pixels on Caltech-256. We use CSIFT/SIFT to extract descriptors. Under a single scale, the scale size is set to 16×16 pixels. The dimension of the traditional SIFT descriptor [16] is 128. About CSIFT descriptors, RGB-SIFT [16] and YCbCr-SIFT [2] are implemented with the dimension of 128×3.
2) Codebook creation. After CSIFT/SIFT descriptors are extracted, a codebook can be created using the K-means clustering method on a randomly selected subset (with size 2×10^6) or on an available set from the whole training dataset. The size of the

Digital Object Identifier 10.1109/MCSE.2018.108164708 1521-9615/$26.00 2018 IEEE

codebook is chosen to be 1024, so as to comparing with other methods;

3) *Localized-soft-assignment coding*. In this paper, CSIFT/SIFT descriptors are encoded by the LSC method using the above constructed codebooks. As searching K-nearest neighbors, we adopt a simple and very efficient K-NN search strategy. The number of neighbors is set to 5 when comparing with other methods;
4) Pooling with *spatial pyramid matching* [3]. Specifically we use two pooling methods, sum pooling and max pooling for comparison. The max-pooling operation is better to compute the final descriptor of each image. In order to incorporate spatial information, it is performed with a 3 levels SPM kernel (1×1, 2×2 and 4×4 sub-regions in the corresponding levels). Each spatial sub-region has the same weight in the "SPM" layer. The pooled features of the sub-regions are concatenated and normalized to form the final descriptor of each image;
5) Classification. A linear SVM classifier is used to train the classifier, since it has shown good performances. In our experiment, we apply a novel ELM learning classifier, which is compared to the linear SVM. We achieve better classification performance as shown below.

B. Assessment of Feature Coding Methods with Color Descriptors on colored Caltech-101 and Caltech-256 Datasets

We use the LSC method with the color SIFT descriptions to obtain better performance on the colored Caltech-101 and Caltech-256 datasets compared to C-SIFT descriptor based LLC method [2] [9]. We list the same experimental setting on the two datasets (colored Caltech-101 and Caltech-256): 1) choosing the three better descriptors (SIFT, YCbCr-SIFT and RGB-SIFT, which are show in paper [9]); 2) the codebook size is fixed to 1024 on Caltech-101 and 4096 on Caltech-256; 3) using 30 images per class for training while leaving the remaining for testing on colored Caltech-101 and 60 on colored Caltech-256. In paper [9], which is part of our experiments. It has been witnessed that the maximum numbers of images per category (30 images on colored Caltech-101 and 60 on colored Caltech-256) for training and the rest for testing can obtain the best performance.

The comparison is summarized in Table I, and it is indicated that our color SIFT descriptors with LSC coding method can outperform the existing techniques, LLC coding method with CSIFT. It shows that YCbCr-SIFT obtains the average classification accuracy of 73.1%, and RGB-SIFT provides the second best average classification accuracy (72.74%)on Caltech-101 dataset. Approximately 1% improvement in classification accuracy can be achieved found by employing YCbCr-SIFT descriptors. On Caltech-256 dataset, YCbCr-SIFT descriptor achieves the average classification accuracy of 44.42%; moreover, RGB-SIFT also provides the second best average classification accuracy (42.48%). Even though, compared to the traditional SIFT descriptors, YCbCr-SIFT descriptor can brought approximately 3% enhancement on the classification accuracy. Hence, it is very significant to adopting YCbCr-SIFT descriptor with LSC leads the competitive performance in image classification tasks.

TABLE I
CLASSIFICATION RATE(%) COMPARISON ON CALTECH-101 AND CALTECH-256

Comparison On Caltech-101		
Feature Coding Methods	CSIFT Descriptors	Classification Accuracy (%)
LSC	SIFT	72.18±0.82
	YCbCr-SIFT	**73.10±0.82**
	RGB-SIFT	72.74±0.94
LLC [2]	SIFT	68.17±0.98
	YCbCr-SIFT	**69.18±1.19**
	RGB-SIFT	68.65±1.33
Comparison On Caltech-256		
Feature Coding Methods	CSIFT Descriptors	Classification Accuracy (%)
LSC	SIFT	41.32±0.45
	YCbCr-SIFT	**44.42±0.27**
	RGB-SIFT	42.48±0.33
LLC [2]	SIFT	37.22±0.35
	YCbCr-SIFT	**41.31±0.27**
	RGB-SIFT	38.71±0.38

C. Assessment of ELM learning method on the original Caltech-101 Dataset

In this experiment, all 102 categories on the challenging original Caltech-101 dataset are used for image classification. As before, we conduct experiments with 30 images per class for training and the rest are used for test. We assess the localized soft-assignment coding enhanced with the ELM learning method (LSC+ELM). We represent the average classification accuracy and the standard deviation (between parentheses) in Table I, and compare the results to those of some famous method. We find that most of the results recorded in the literature on Caltech-101 rely on SIFT features only. Consequently, we also report results with SIFT features in our experiment.

We divide the table into two sections as in paper [1]. The top one is our proposed method (the LSC

Digital Object Identifier 10.1109/MCSE.2018.108164708 1521-9615/$26.00 2018 IEEE

TABLE II
CLASSIFICATION RATE(%) COMPARISON ON CALTECH-101 DATA SET

Training images	Classification Accuracy (%)
Ours(LSC+ELM)	**76.97±0.41**
LSC (Ours)	72.63±1.13
LSC [1]	74.21±0.81
LLC [8]	73.44± –
Soft-assignment coding [18]	64.1±1.2
Soft-assignment coding [17]	69.0±0.8
Sparse coding [4]	73.2±0.55

TABLE IV
CLASSIFICATION RATE(%) COMPARISON ON CALTECH-101 DATASETS

Training images	Classification Accuracy(%)
Ours(SIFT+LSC+ELM)	76.58±0.40
Ours(RGB-SIFT+LSC+ELM)	76.72±0.46
Ours(YCbCr-SIFT+LSC+ELM)	**78.0±0.47**
LSC [1]	74.21±0.81
LLC [2]	68.17±0.98
LLC(RGB-SIFT+LLC) [2]	68.65±1013
LLC(YCbCr-SIFT+LLC) [2]	69.18±1.19

combined with the ELM method), and the second is the original localized soft-assignment coding applied by ourselves. The original LSC (ours) is used the same method in paper [1] and implemented in our experiment conditions. As Table II shows, our approach is better than the original LSC (ours). In the bottom part, we present several versions of the existing coding approaches in the literature. We list the localized soft-assignment coding in [1], which is one of the most popular feature coding schemes. According to the experiment results, although we use the similar experimental environment as the paper [1], the result of our LSC in top is 72.63%. It is still a little lower than them in the paper [1]. However, our proposed method (LSC+ELM) shows the best performance 76.97% in Table I. Additionally, as LLC is one of the state-of-art coding methods, our approach (LSC+ELM) reaches comparable performance or higher performance than LLC. We provide two kinds of soft-assignment, one of them uses the sum-pooling in [17] and the other takes max-pooling in [18]. The sparse coding test is provided too. Considerably, this employment in [17] [18] is more comparable to our approach than the traditional soft-assignment. In our evaluation, our method presents better performance than different versions of coding methods, and it outperforms them with more than 3% improvement on average classification accuracy.

D. Assessment of the new final representation framework on the colored Caltech-101 Dataset

Form the previous discussions, we find that both RGB-SIFT and YCbCr-SIFT outperform the traditional SIFT. Therefore, we only take these three color descriptions with LSC and ELM classifier in our experiment. The results are represented in Table III, and again YCbCr-SIFT descriptor achieves the best performance. For example, while the number of training images is 30, YCbCr-SIFT obtains the best average classification accuracy at 78.0%; RGB-SIFT provides the second best average classification accuracy (76.72%). As shown from Table IV, the experimental results are extremely impressive: under the same cases, our proposed new framework shows higher Classification performance by more than 4% over in paper [1]; In addition, approximately 8% or 9% improvement than the existing technique (LLC with CSIFT), which is the color SIFT descriptions based Locality-constrained Linear Coding and list in the literature [2].

V. CONCLUSION

We integrated CSIFT descriptors and ELM learning method to improve the LSC based image classification system. With various settings of the parameters, two kinds of CSIFT descriptors are implemented and evaluated in our experiment. About the CSIFT has been exploit in my conference paper [9], which is parts of work in our paper. Among the experiment results, we noticed that YCbCr-SIFT descriptor achieves the most stable and accurate image classification performance. Compared to the highest average classification accuracy achieved by using traditional SIFT descriptors, YCbCr-SIFT descriptor acquired approximately 1% increase on the Caltech-101 dataset and approximately 3% increase on the Caltech-256 dataset.

For invariant or discriminatory object recognition, we propose to use the ELM method to enhance the traditional BoF framework. The main contribution of this paper is that we used the ELM learning algorithm to build a novel image classification system. In the experimental result we show that after we boost the localized soft-assignment coding with ELM learning approach, it gives a better performance. Compared to the different versions of coding methods, our method (ELM+LSC) outperforms them with more than 3% or 4% improvement in average classification accuracy.

Finally, we also propose the new final representation framework in order to enhance the classification results. In fact, our new proposed framework shows higher Classification performance by more than 3% or 4% on the Caltech-101 dataset. We found that variations of CSIFT descriptors can be used in many image classification problems. The ELM learning method is also

Digital Object Identifier 10.1109/MCSE.2018.108164708 1521-9615/$26.00 2018 IEEE

TABLE III
CLASSIFICATION RATE(%) COMPARISON ON CALTECH-101 DATASRTS

Training images	5	10	15	20	25	30
RGB-SIFT+ LSC+ELM	59.30±0.83	66.90±0.92	70.89±0.76	73.49±0.45	75.25±0.23	76.72±0.46
YCbCr-SIFT+ LSC+ELM	**60.23±0.59**	**67.86±0.72**	**71.58±0.65**	**73.94±0.45**	**76.30±0.40**	**78.0±0.47**
SIFT+ LSC+ELM	59.25±0.74	67.17±0.94	70.83±0.76	73.39±0.36	75.23±0.43	76.58±0.40

very useful in object recognition. Combining different CSIFT descriptors, ELM and deep learning method are our future research.

REFERENCES

[1] L. Liu, L. Wang, and X. Liu, "In defense of soft-assignment coding," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2486–2493.

[2] J. Chen, Q. Li, Q. Peng, and K. H. Wong, "Csift based locality-constrained linear coding for image classification," *Pattern Analysis and Applications*, vol. 18, no. 2, pp. 441–450, 2015.

[3] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2. IEEE, 2006, pp. 2169–2178.

[4] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1794–1801.

[5] Z. Wang, J. Feng, S. Yan, and H. Xi, "Linear distance coding for image classification," *Image Processing, IEEE Transactions on*, vol. 22, no. 2, pp. 537–548, 2013.

[6] J. C. van Gemert, C. J. Veenman, A. W. Smeulders, and J.-M. Geusebroek, "Visual word ambiguity," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 7, pp. 1271–1283, 2010.

[7] K. Yu, T. Zhang, and Y. Gong, "Nonlinear learning using local coordinate coding," *Advances in Neural Information Processing Systems*, vol. 22, pp. 2223–2231, 2009.

[8] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3360–3367.

[9] Q. Li, J. Chen, Q. Peng, and X. Wu, "Application of localized soft-assignment coding and csift in image classification," in *Proceedings of International Conference on Internet Multimedia Computing and Service*. ACM, 2014, p. 246.

[10] B. He, D. Xu, R. Nian, M. van Heeswijk, Q. Yu, Y. Miche, and A. Lendasse, "Fast face recognition via sparse coding and extreme learning machine," *Cognitive Computation*, vol. 6, no. 2, pp. 264–277, 2014.

[11] H. Yang, J. Yi, J. Zhao, and Z. Dong, "Extreme learning machine based genetic algorithm and its application in power system economic dispatch," *Neurocomputing*, vol. 102, pp. 154–162, 2013.

[12] A. Shabou and H. LeBorgne, "Locality-constrained and spatially regularized coding for scene categorization," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3618–3625.

[13] S. McCann and D. G. Lowe, "Spatially local coding for object recognition," in *Computer Vision–ACCV 2012*. Springer, 2013, pp. 204–217.

[14] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 42, no. 2, pp. 513–529, 2012.

[15] G.-B. Huang, M.-B. Li, L. Chen, and C.-K. Siew, "Incremental extreme learning machine with fully complex hidden nodes," *Neurocomputing*, vol. 71, no. 4, pp. 576–583, 2008.

[16] K. E. van de Sande, T. Gevers, and C. G. Snoek, "Evaluating color descriptors for object and scene recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 9, pp. 1582–1596, 2010.

[17] J. C. van Gemert, J.-M. Geusebroek, C. J. Veenman, and A. W. Smeulders, "Kernel codebooks for scene categorization," in *Computer Vision–ECCV 2008*. Springer, 2008, pp. 696–709.

[18] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2559–2566.

Qing Li received the B. E. degrees in computer science and received the M.S. in traffic information engineering and control from Southwest Jiaotong University, Chengdu, China, in 2009 and 2013. She is currently working towards the Ph.D. degree in computer application technology, Southwest Jiaotong University, Chengdu, China. Her research interests include computer vision, image classification, and machine learning.

Qiang Peng received the B.E in automation control from Xi'an Jiaotong University, Xi'an, China, and the M.Eng. in computer application technology and Ph.D. degrees in traffic information and control engineering from Southewest University, Chengdu, China, in 1984, 1987, and 2004, respectively. He is currently a Professor at the School of Information Science and Technology, Southwest Jiaotong University, Chengdu, China. From 2003 to 2004, he was with School of Electrical and Electronic Engineering, Singapore Polytechnic, where he served as a visiting research scholar on algorithms for embedded system and intelligent robots. His research interests include digital video compression and transmission, image/graphics processing, traffic information detection and simulation, virtual reality technology, multimedia systems and applications.

Digital Object Identifier 10.1109/MCSE.2018.108164708 1521-9615/$26.00 2018 IEEE

Junzhou Chen received his Ph.D. degree from Department of Computer Science & Engineering of the Chinese University of Hong Kong. He also received a bachelor degree from Sichuan University. He is currently an associate professor at School of Information Science and Technology, Southwest Jiaotong University, Chengdu, China. His research interests includes: Computer Vision, Pattern Recognition, Machine Learning, and Image/Video Processing.

Chuan Yan received the B. E. degrees in communication engineering from Southwest Jiaotong University, Chengdu, China, in 2007. He is currently working towards the Ph.D. degree in computer application technology, Southwest Jiaotong University, Chengdu, China. His research interests include perceptual video coding, deep learning and high efficiency video coding technologies.

Digital Object Identifier 10.1109/MCSE.2018.108164708 1521-9615/$26.00 2018 IEEE