Multimed Tools Appl (2017) 76:23059–23075
DOI 10.1007/s11042-016-4316-z

Land-use scene classification: a comparative study on bag of visual word framework

Mana Shahriari[1] · Robert Bergevin[1]

Received: 2 June 2016 / Revised: 14 December 2016 / Accepted: 27 December 2016 /
Published online: 11 January 2017

Abstract With successful launch of high spatial resolution (HSR) sensors, highly detailed spatial information is provided for remote sensing research. This improvement has allowed researchers to monitor environmental changes on a small spatial scale. However traditional pixel-based classification approaches are not able to interpret high spatial resolution remote sensing imagery effectively. Bag of visual words (BoVW) framework, on the other hand, is becoming one of the most popular approaches to validate the performance of remote sensing image datasets. While pixel-based approaches may not fully describe very high-resolution remote sensing images, BoVW model is narrowing the gap between low-level features and high-level semantic features by generating an intermediate description of image features. This paper presents a comparative study to evaluate the potential of using different coding approaches of BoVW model to solve the land-use scene classification problem. Initially, this work summarizes different configurations of BoVW framework in coding and clustering. Later, we perform an extensive evaluation of BoVW on land-use scene classification and retrieval. Finally we draw several conclusions regarding different coding strategies of BoVW, codebook size and number of training images. The approach is validated on two commonly used datasets in remote sensing, UC Merced a 21-class land-use dataset and RSDataset a 19-class satellite scene dataset.

Keywords Remote sensing · Land-use scene classification · Bag of visual words · Spatial pyramid · SIFT · Encoding · Clustering

✉ Mana Shahriari
mana.shahriari.1@ulaval.ca

Robert Bergevin
robert.bergevin@gel.ulaval.ca

[1] Computer Vision and Systems Laboratory, Laval University, Quebec, QC, Canada

1 Introduction

Automated land-use scene classification has become widely desirable due to the amount of high-resolution remote sensing image data. Remote-sensed airborne and satellite imaging aims at predicting the content labels that globally describe a given image, i.e. land-use scene labels. It is used in many fields such as agriculture, geography, military, humanitarian applications, and many other applications for analyzing and managing natural resources. There has been a great deal of effort for developing intelligent databases for effective and efficient processing of high-resolution remote-sensed images. With the development of modern technologies, the resolution has been improving in remote sensing images and more detailed spatial information can be obtained. Due to this improvement and in addition to the widely used spectral information [30], novel computational approaches are constantly required.

Land-use scene classification has been explored from a variety of angles in the literature in the last decades. Generally speaking, the process of recognition starts by extracting a set of features from training data and follows by deriving a classifier to label test data. Thanks to developments in acquisition of high-resolution remote sensing images, extraction of color, texture, shape and object information has become possible [28, 29, 40, 41]. Since high-spatial resolution images contain rich textural information, the approaches which capture the texture information are widely adopted in this domain. With the success of local binary patterns (LBP) [22] as a texture descriptor, this model has been used in many land-use scene classification tasks [4, 5, 27, 45]. In [1], a semantic modeling based method is developed to fill the gap between low-level features and high-level user expectations and scenarios. A Bayesian framework for a visual grammar is employed in their framework to reduce this gap. Chen et al. [5] incorporates multi-orientation Gabor filters to capture the global texture information from an input image, and LBP to describe it locally. An improved Gaussian Markov random field (GMRF) model is used in [45] to extract texture features from high-spatial resolution images. While the effectiveness of such texture features has been verified, the classification procedure is later completed by combining spatial and spectral features.

However, despite being in the literature for some time, the ability to predict semantic category from pixel level (low-level features) is a hard task. The challenge is due to high variability of image appearance, i.e. variations in spatial position, illumination, and scale. Additionally, pixel-based image classification approaches may suffer the increase of within-class spectral variation with improved spatial resolution [41]. Thus, in a different direction, some other researches use region-based and object-based features to implement land-use scene classifiers [3, 6, 29, 40]. These approaches are built upon image segmentation and unlike single-pixel approaches, spatial information of image regions (i.e. segments) can be modeled [3]. Region-based approaches highly depend on a good segmentation algorithm, and cannot usually capture the complex semantic information due to the semantic phenomena known as ‘synonymy’ and ‘homonymy’ [3, 47].

Local features, on the other hand, are becoming more and more popular since they can provide robustness to rotation, scale changes, and occlusion. In addition, local features can bridge the gap between low-level features and the high-level semantics by building a mid-level feature representation of image regions across image patches. One of the most popular local feature frameworks is the Bag-of-Visual-Words (BoVW) [7, 32], where an image is represented by the histogram of occurrences of vector-quantized descriptors. It is tailored to handle scale and rotation variance, it provides a concise representation of an image, and it has shown decent performance in whole-image categorization tasks [11, 36] including for remote sensing image datasets [12, 25, 43]. In 2010, Yang and Newsam [39] investigated the traditional BoVW approach for land-use scene classification in HSR imagery. Overall, their work does not show an improvement over the best standard approaches, however BoVW represents a robust alternative for certain classes. To improve the performance of the basic BoVW framework, [6, 41, 42, 44] propose adding spatial and context information to the local features. [44] presents a concentric circle-based spatial-rotation-invariant representation strategy for describing spatial information, while [6] propose a pyramid-of-spatial-relatons (PSR) model to capture both absolute and relative spatial relationships of local features. [41, 42] on the other hand combine an object-based mid-level representation method with BoVW for an improved semantic classification. [47, 48] corporate both local and global spatial features for HSR image scene classification. To capture the local features of land-use scene images, both papers use BoVW framework. For global feature extraction [48] propose multi-scale local binary patterns (MS-CLBP), whereas in [47] the shape-based invariant texture index is designed as the global texture feature.

A BoVW framework consist of the following steps (see Fig. 1): (i) key-point extraction, which samples local areas of images and outputs image patches; (ii) key-point representation, where image patches are described via statistical analysis approaches; (iii) codebook generation, which aims at providing a compact representation of local descriptors; (iv) feature encoding, which codes local descriptors based on the codebook they induce; (v) feature pooling, where the final image representation which integrates all coding responses into one vector, i.e. the feature vector is produced; and (vi) classification, obtained from feature vectors using a support vector machine, for instance.

The main contribution of the work reported in this paper is investigating in details different coding and pooling strategies of the BoVW framework. To this end, we carry out a comparative analysis of the BoVW model with different configurations on two commonly used datasets in remote sensing. We draw several conclusions on these datasets when comparing different coding representations of this model. Furthermore, the effect of dictionary size and the number of training images in respect to the coding approaches is studied.

The rest of the paper is organized as follows: after reviewing the steps of a BoVW framework in Section 2, the details on setting up the experiments are given in Section 3. Analysis of results and discussion are coming in Section 4. Finally conclusions appear in Section 5.

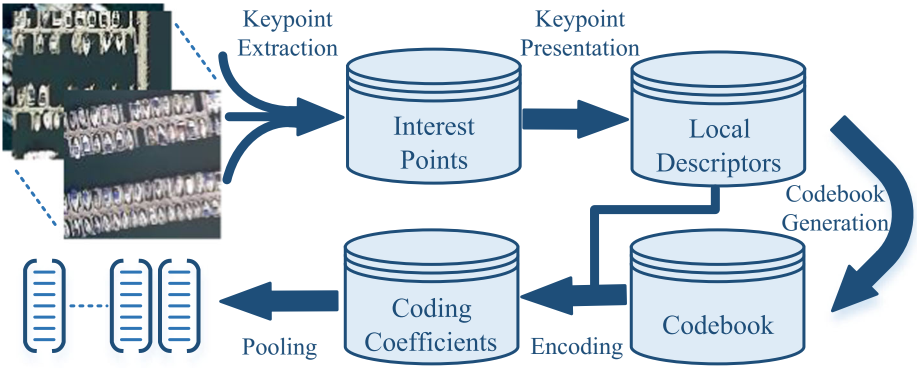

Fig. 1 The general scheme of a bag of visual words (BoVW) framework. It starts by extracting several keypoints (patches) from an input image, and follows by describing and clustering them into the codebook. The final image-level representation is obtained by pooling the coding coefficients of encoded local descriptors. The idea is adopted from [13]

2 Bag of visual words

To compute a global image representation from a large set of local descriptors, various approaches can be taken. The remainder of this section will review the most commonly used methods for each step in a BoVW framework. However, the focus is given to reviewing codebook generation (step (iii)) and feature encoding (step (iv)) in particular. These two steps are considered to be the most important ones for image representation since they have a great impact on the classification performance [11, 17].

2.1 Key-point extraction – step (i)

In a BoVW model, key-points or interest points are locally sampled image patches. These patches are meant to be the most informative regions of the image, and the best candidates for image classification. They can be sampled either sparsely or densely. The notion of sparse key-point extraction [19, 34] (also known as saliency-based sampling) has evolved from edge, corner and blob detection approaches. Dense features on the other hand are sampled on a regular grid and can provide a good coverage of the entire image [33]. Dense features have shown to be an essential component of many state-of-the-art classification approaches [16, 26, 31]. Hu et al. [12], investigate and quantitatively compare different sampling strategies that can be used for scene classification of High resolution remote sensing images.

2.2 Key-point representation – step (ii)

Patch representation is the process of describing the pixels of local image patches (interest points) statistically. Patch descriptor approaches can compute local features which are invariant to image scale and rotation, and provide robustness to illumination changes, noise, and changes in viewpoint.

Let N denote a set of N image patches, then the vector of local descriptors X can be defined as: $X = [x_1, x_2, \ldots, x_N]$ where $x_i \in R^D, i = 1, 2, \ldots, N$ and D is the dimensionality of local features computed by an image descriptor. In the domain of scene understanding, object classification, and image retrieval, the most widely used patch descriptors are scale-invariant feature transform (SIFT) [18], histogram of oriented gradients (HOG) [9] and local binary patterns (LBP) [22]. SIFT descriptor compute the orientation and gradient of the key points in gray-level information, and exhibits powerful description capability in land-use scene classification [12, 44, 47, 48].

2.3 Codebook generation – step (iii)

The codebook is a concise representation of local descriptors and seeks two main goals: (1) avoiding redundancy, (2) adding robustness to scene layout by providing invariant features. A codebook can be seen as a collection of basic patterns (visual words / codewords) used to reconstruct local descriptors. The collection of visual words which is called the vocabulary of visual words or the codebook is commonly generated through clustering in a supervised or an unsupervised manner [13].

Given the vector of local descriptors X, any clustering approach seeks the K basis vectors c_j (or visual words) where $K \ll N$. The idea is to reconstruct X, using a set of basis vectors (i.e. visual codebook) $C = [c_1, c_2, \ldots, c_K] \in R^D$ and the coding coefficients. The coding coefficient component of x_i with respect to the visual word c_j is called u_{ij}.

2.3.1 Hard assignment

Hard-assignment is considered the simplest and the most common clustering approach in the literature. Usually k-means is adopted to quantize descriptors into the visual vocabulary such that the cumulative approximation error is minimized:

$$u_{ij} = \begin{cases} 1; & \text{if } i = \underset{k=1,2,\ldots,K}{\operatorname{argmin}} \ (\|x - c_k\|_2^2) \\ 0; & \text{otherwise} \end{cases} \tag{1}$$

where $\|.\|_2$ denotes the Euclidean distance between the descriptor vector x and the visual words c_k. As (1) implies, k-means assigns a local feature to its closest visual word. However this type of assignment can cause severe information loss specially for features located on the boundaries of several visual words.

2.3.2 Soft assignment

Soft assignment [24] aims at minimizing the quantization error by assigning each descriptor to more than one codeword. This type of assignment is the weighted assignment of the descriptor x_i to the jth visual word c_j based on its distance / similarity. The coding coefficient u_{ij} is controlled by the smoothing factor β and represents the degree of membership of x_i to c_j:

$$u_{ij} = \frac{\exp(-\beta\|x_i - c_j\|_2^2)}{\sum_{k=1}^{K} \exp(-\beta\|x_i - c_k\|_2^2} \tag{2}$$

The denominator is the normalization factor. However soft assignment results in "dense code vectors, which is undesirable, among other reasons, because it leads to ambiguities due to the superposition of the components in the pooling step" [2].

2.3.3 GMM clustering

One way to mitigate the issue of ambiguities in soft assignment clustering is to use GMM clustering. In Gaussian mixture model (GMM) clustering, the probability density distribution of features is defined by a collection of Gaussian distributions. GMM can be thought of as a soft visual vocabulary. The coding coefficient u_{ij} is:

$$u_{ij} = \frac{p(x_i|\mu_j, \Sigma_j)\omega_j}{\sum_{k=1}^{K} p(x_i|\mu_k, \Sigma_k)\omega_k} \tag{3}$$

Each GMM $p(x|\theta)$ represents a cluster of data points (a set of descriptors), and is fully defined by its vector of parameters, i.e. the weight ω_k, mean μ_k, and the covariance matrix Σ_k.

$$\begin{aligned} p(x|\theta) &= \sum_{k=1}^{K} p(x|\mu_k, \Sigma_k)\omega_k \text{ where} \\ p(x|\mu_k, \Sigma_k) &= \frac{1}{\sqrt{(2\pi)^D|\Sigma_k|}} \exp(-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k)) \end{aligned} \tag{4}$$

The vector of parameters $\theta = (\omega_1, \mu_1, \Sigma_1, \ldots, \omega_K, \mu_K, \Sigma_K)$ is learnt iteratively by the Expectation Maximization (EM) algorithm [20].

2.4 Feature encoding – step (iv)

In general, the purpose of encoding is to statistically summarize a number of local feature descriptors given the codewords. In addition to that, the most recent encoding approaches aim at reducing the loss of information introduced while the codebook is formed (Section 2.1). The coding can be seen as a activation function for the codebook, which is activated for each of the codewords according to the local descriptor [2].

2.4.1 Histogram coding

In the classical BoVW representation, each bin of the histogram (codeword) is activated only by its closest descriptor (1). Given a set of descriptors $x_1, \ldots, x_N$, the frequency of each bin reflects how many times each bin (visual word) is activated by each x_i. The frequency histogram based image representation is considered the baseline encoding approach. However it suffers from instabilities when the descriptors are located on the boundaries of several codewords (due to the quantization error of hard assignment).

2.4.2 Kernel codebook coding

To enhance the accuracy of probability density estimation, this type of encoding is suggested by [20]. Each descriptor activates multiple codewords, that is to say descriptors are assigned to the codewords in a soft manner. Equation (2) can be rewritten as:

$$p(c_j|x_i) = \frac{1}{Z} exp(s(x_i, c_j)) \text{ where } s(x_i, c_j) = -\beta \|x_i - c_j\|_2^2 \tag{5}$$

showing the probability of the local descriptor x_i belonging to the codeword c_j. Z, the normalization factor, ensures that $\sum_{k=1}^{K} P(c_k|x_i) = 1$. In the original form of soft coding, all the K visual words are used to compute the coding coefficients.

2.4.3 Fisher coding

In Fisher coding, the descriptors are encoded using a kernel function derived from a generative probability model $p(x|\theta)$. This can be done by fitting a parametric generative model θ (e.g. GMM) to descriptors (4). Later each descriptor x_i is represented by the gradient of the log-likelihood with respect to GMM parameters [14]. If the covariance matrices, Σ_k are assumed to be diagonal [23], the gradients are considered with respect to the mean and standard deviation.

For each Gaussian mode k, consider the mean and covariance deviation vectors as follows:

$$\Phi_{\mu,k} = \frac{1}{N\sqrt{\omega_k}} \sum_{i=1}^{N} \alpha_i(k) \big(\frac{x_i - \mu_k}{\Sigma_k}\big) \tag{6}$$

and

$$\Phi_{\Sigma,k} = \frac{1}{N\sqrt{2\omega_k}} \sum_{i=1}^{N} \alpha_i(k) \big[\big(\frac{x_i - \mu_k}{\Sigma_k}\big)^2 - 1\big] \tag{7}$$

where $\alpha_i(k)$ is the soft assignment weight of descriptor x_i to Gaussian k. This representation captures the average first and second order differences between the descriptor and each of the GMM centers [31]. The final gradient vector, i.e. Fisher vector Φ, is obtained by aggregating $\Phi_{\mu,k}$ and $\Phi_{\Sigma,k}$ over all K GMMs, and is $2KD$-dimensional. The Fisher encoding can be further improved (Improved Fisher Vector – IFV) through two normalization steps: power and L_2 normalization [23].

2.4.4 Sparse coding

Sparse coding (SC) [38], along with other coding approaches intends to ameliorate the quantization loss of vector quantization, e.g. k-means. The core idea is to reconstruct the local descriptor x_i by a linear combination of a set of sparse codewords. Generally, this can be done by solving a least-square-based optimization problem.

$$\underset{U}{\operatorname{argmin}} \sum_{i=1}^{N} \|x_i - Cu_i\|^2 + \lambda \|u_i\|_{l^1} \tag{8}$$

The first term (the least-square term), pursues accurate reconstruction. Whereas the second term (sparsity regularization term) l^1 norm of c_i ensures the sparsity is met.

2.4.5 Locality constrained linear coding

Locality-constrained Linear coding (LLC) [36] can be seen as a variant of sparse coding. However, instead of the sparsity constraint, LLC incorporates a locality constraint. The LLC encoding of the local descriptor x_i is of size K, having M non-zero components closest to the visual word x_i (where $M \ll K$). Mathematically LLC can be formulated as:

$$\underset{U}{\operatorname{argmin}} \sum_{i=1}^{N} \|x_i - Cu_i\|^2 + \lambda \|d_i \odot c_i\|^2 \text{ s.t. } 1^T u_i = 1 \forall i \tag{9}$$

where $\odot$ denotes the element-wise multiplication. $d_i = \exp(dist(x_i, C)/\sigma)$ and $dist(x_i, C) = [dist(x_i, c_1), \ldots, dist(x_i, c_M)]^T$ is the Euclidean distance between x_i and c_j. Finally σ is a weighting parameter, controlling the decay speed for the locality adapter.

2.4.6 Super-vector coding

Zhou et al. [46] propose a coding scheme, that is the super-vector coding (SVC), which shares some similarities with IFV and LCC. The idea is to estimate the feature manifold by deriving a non-linear mapping function (i.e. $f(x) \approx \omega^T \Phi(x)$). SVC typically uses the closest codewords of a feature (hard assignment), however there is another variant based on soft assignment. It uses the first order statistics between local descriptors X and the codebook C and adds it to the adaptive representation histogram for better reconstruction:

$$\Phi(x) = [s\gamma_c(x), \gamma_c(x)(x-c)^T]^T_{c \in C} \tag{10}$$

where $\gamma_c(x)$ is essentially the coding coefficients of codeword c_j for the local descriptor x_i and s is a non-negative constant. The resulting vector is a highly sparse representation with the dimensionality of $K(D+1)$.

2.5 Pooling – step (v)

Pooling is the final step of a BoVW framework, where the idea is to obtain an image-level representation from the coding coefficients u_{ij} of local features x_i in the image. Suppose that we have responses of the visual word c_j for all the local descriptors $x_i, i = 1, \ldots, N$. Then, pooling can be seen as a function summarizing visual word c_j for local descriptors, and "may be interpreted as the aggregation of the activations of that codeword" [2]. Two of the most common pooling functions are sum (or average) and max pooling. Given the responses for the j^{th} visual word c_j, the i^{th} component of the final image-level representation: for sum pooling it is $\sum_{i=1}^{N} u_{ij}$; for average pooling it is $\frac{1}{N} \sum_{i=1}^{N} u_{ij}$; and for max pooling is: $max_i u_{ij}$. Sum / Average pooling preserves the average response and is widely used in traditional BoVW. However, max pooling preserves the maximum response, and is often preferred for sparse and soft coding.

3 Performance evaluation

3.1 Coding approaches

The following five coding approaches are chosen to extensively evaluate the performance of two common land-use scene datasets, that is UC Merced Land Use Dataset [39] and High-Resolution Satellite Scene Dataset [8, 37]. It is noteworthy that both datasets have high spatial resolution (HSR) remote sensing images only and thus optical RGB images are used for all the experiments. In addition to that, throughout all the experiments, only the luminance channel is used.

1. *Histogram Coding* with hard quantization is selected as the baseline coding method for the BoVW framework;
2. *Kernel Codebook Coding* or soft-assignment coding is selected as the representative of a soft quantization approach;
3. *Locality-constrained Linear Coding – LLC* is selected as a good representative of sparse coding scheme considering both the accuracy and the computational cost;
4. *Improved Fisher Vector – IFV* is selected since it has shown to be a powerful coding approach for image-level presentation;
5. *Super-Vector Coding* is selected as a simple extension of histogram coding that is similar to both IFV and similar to LCC but in different ways.

3.2 UC Merced land use dataset – UCMerced

The methods are first evaluated using a large ground truth image dataset of 21 land-use classes [39]. There are 100 images for each of the following classes: agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium density residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis courts. Each image contains 256 × 256 pixels, and was manually cropped from large images from the USGS National Map Urban Area Imagery collection for various urban areas around the USA [39]. Following the common benchmarking procedure in [39], 80 images from each land-use are chosen for training and the rest are used for evaluating the performance. The experiments are then repeated 10 times, with different randomly selected training and test images. The final result is reported as the mean and standard deviation of the results from the individual runs. Sample images of this dataset appear in Fig. 2.

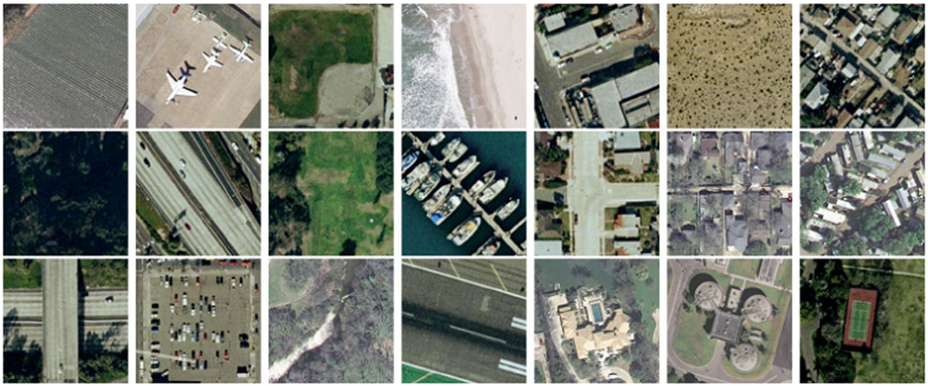

Fig. 2 Sample images from each class of a 21 land-use dataset. One example per each class is shown above. From *left* to *right*, *top* to *bottom* the classes are: agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium density residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis courts

3.3 High-resolution satellite scene dataset – RSDataset

The second dataset used is a 19-class satellite scene dataset including airport, bridge, commercial, desert, farmland, football field, forest, meadow, mountain, park, parking, pond, port, railway station, river, viaduct, commercial area, industrial area, and residential area [8, 37]. There are at least 50 images of size 600 × 600 per each class. In order to facilitate a fair comparison, the same experimental setup as suggested in [4] is followed. To this end, 30 random images per class are chosen to train the models, and the remaining are used for testing. Like with previous dataset, the experiments are repeated 10 times. The final result is reported as the mean and standard deviation of the results from the individual runs. Some sample images of this dataset is given in Fig. 3.

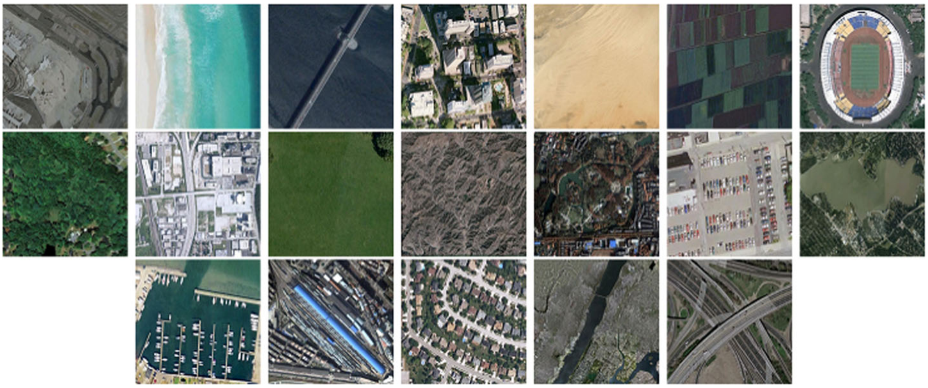

Fig. 3 Sample images from each class of a 19 satellite scene dataset. One example per each class is shown above. From *left* to *right*, *top* to *bottom* the classes are: airport, beach, bridge, commercial area, desert, farmland, football field, forest, industrial area, meadow, mountain, park, parking, pond, port, railway station, residential area, river, viaduct

3.4 Experimental setup

In this study, all the descriptors are extracted over a dense grid of pixels, rather than a sparse set of interest points, as suggested by [16, 26, 31]. The local descriptors $X = [x_1, \ldots, x_N]$ are the scale-invariant feature transform (SIFT) descriptors [18]. Local descriptors are extracted densely in scale and space, on spatial grids equal to 8, 12, 16, 20 and 24 pixels, with the step size set to 2 pixels using the publicly available VLFeat toolbox [35]. As suggested by the authors in the original literature, max pooling is used for LLC encoding, weighted average pooling for super-vector coding, and sum/average for all the other approaches. The final image-level representation is passed through signed square-rooting and is L_2 normalized before being sent to the linear SVM for training and classification (LIBLINEAR package [10]). The number of visual words varies between 16 and 16384 for both datasets.

3.5 Spatial pyramid

The BoVW framework provides a flexible visual layout of images by presenting them as an orderless collection of local features. This results in a holistic representation of an image, where the spatial information of features is lost and ignored. To tackle this issue, we incorporate spatial information using the spatial pyramid matching scheme from [16]. The spatial pyramid matching (SPM) has shown to be successful in object and scene recognition. The basic SPM approach starts by partitioning an image into a sequence of increasingly fine sub-regions. The histograms are then computed for each grid and are stacked together after being weighted and normalized according to their size. This concept can be easily extended to any of the BoVW methods by encoding the regions and stacking the feature vectors. In [15] another variant of the basic SPM [16] is suggested, where each spatial region is normalized individually prior to stacking. The final feature vector is l^2 normalized before being sent to the SVM classifier. Like [16], the spatial regions for both datasets are set to be 1×1, 2×2, and 3×1 grids, and a total of 8 regions.

4 Analysis of results

Table 1 presents the overall performance of each of the coding approaches discussed in Section 2 and Section 3 and the comparison of the results with the state-of-the-art. Note that the results in this table are quoted directly from the papers, where the best overall performance is chosen for comparison. In the remainder of the paper we are going to call histogram quantization coding HQ, kernel codebook coding KC, locality-constrained linear coding LLC, super vector coding SVC, and improved Fisher vector IFV.

Based on the result we got in (see Table 1), the following observations can be made:

1. As expected HQ is the baseline coding approach for BoVW framework and its result stands well below other coding approaches (except for KC). The improvement with the other coding strategies (LLC, SVC, and IFV) shows the evolution of coding approaches over the time.
2. For both datasets HQ and KC produce almost the same results. Although soft-assignment is used instead of hard-assignment, almost no improvement in classification is observed. The inferior performance of KC is due to the ambiguities introduced in the

Table 1 Performance evaluation of the two commonly used land-use datasets; UCMerced [39] and RSDataset [8, 37]. Following the common benchmarking procedure, 80 training images are used for UCMerced and 30 training images for RSDataset. The comparison between the coding methods discussed in this paper as well as the state-of-the-art performances is provided

Methods	**UCMerced**	**RSDataset**
Color RGB Histograms [39]	76.7%	–
Homogeneous Texture (Gabor Filters) [39]	76.9%	–
Spatial Extensions for BoVW [39]	77.71%	–
Color HLS Histograms [39]	81.19%	–
Improved Correlatons [25]	82.13%	–
Completed Local Binary Pattern [4]	85.48%	92.37%
Concentric Circle-Structured BoVW [44]	86.64± 0.81%	–
2-D Wavelet Decomposition BoVW [43]	87.38±1.27%	–
Pyramid of Spatial Relatons [6]	89.1%	–
Gabor-Filtering Completed LBP [5]	90.0±2.1%	91.0±1.5%
Local Global Fusion (LGF) [48]	95.48%	95.26%
Local–Global Feature BoVW [47]	96.88±1.32	–
Histogram Quantization Coding (HQ)	79.62±2.34%	80.81±2.96 %
Kernel Codebook Coding (KC)	79.14±1.41%	81.95±2.27%
Locality-constrained Linear Coding (LLC)	87.12±1.28%	87.16±0.94%
Super Vector Coding (SVC)	90.90±1.62%	91.73±0.82%
Improved Fisher Vector (IFV)	95.07±0.93%	95.81±0.51%

pooling step. As proposed by Liu et al. [17] the accuracy of KC can be improved when localized soft-assignment is combined with mix-order max-pooling.

3. When local descriptors are clustered into codewords, some part of information is lost. LLC and sparse coding in general employ a least-squares based optimization to lessen the information loss. As the results suggest, we can see the improvement over HQ and KC when LLC is applied.
4. Overall, the classification accuracy of IFV outperforms all other coding schemes. Comparing IFV results with the state-of-the-art (in particular with Local Global Fusion (LGF) [48]) implies that this approach perform just as well as the best land-use scene classification approaches.
5. SVC outperforms LLC, however its performance cannot reach the classification accuracy of IFV. Generally speaking, SVC runs faster than IFV and combines computational efficiency and a good and reasonable classification performance.

Next, the influence of vocabulary size on performance of the five coding schemes is evaluated (see Fig. 4):

1. As seen in Fig. 4, the overall tendency is that a larger number of visual words leads to a higher accuracy. In the case of HQ, KC, and LLC and for both datasets, the performance degrades dramatically when a smaller codebook size is chosen.
2. For all coding approaches and after a certain codebook size, an over-fitting effect is noticed. This varies for different coding approaches and for different datasets. For instance for both datasets and for HQ and VC coding the saturation occurs at about

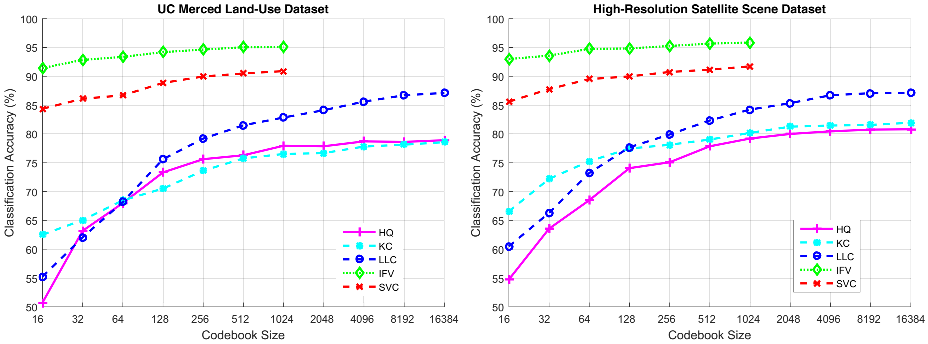

Fig. 4 Comparing the performance of on UCMerced (*left*) and RSDataset (*right*) when varying the size of codebook. The size of codebook varies from 16 to 16384 for both datasets. Following the common benchmarking procedure, 80 training images are used for UCMerced and 30 training images for RSDataset

1024 visual words. However LLC on UC Merced is not saturated even at 16384 visual words. For this case, the performance can increase even further if the codebook size is increased to 32K (32768 ≈32K).

3. In the case of SVC and IFV, the over-fitting occurs for a smaller vocabulary size. In both cases the final image-level representation is very high dimensional. Thus choosing an optimal number of codewords plays an important role from the computational efficiency point of view. Perronnin [23] suggested the size of codebook for IFV to be set to 256. This value is validated in different experiments [13, 15, 23, 31] and provides a good compromise between computational cost and classification accuracy. A larger codebook size typically increases the accuracy by only a few decimals at a higher computational cost. For both IFV and SVC cases, the codebook size is not increased further than 1024, due to the saturation, and the computational complexity of the approach.

Furthermore, the effect of the number of training samples is investigated and presented in Table 2. We change the number of training images from 10 to 80 for UCMereced and from

Table 2 Comparing the classification accuracy of UCMerced (*top*) and RSDataset (*bottom*) when varying the number of training images. The codebook size is fixed for both datasets and is set to 16384 for all five coding strategies

UCMerced	**10**	**20**	**40**	**80**
Histogram Quantization Coding	51.66%	66.77%	71.25%	79.62%
Kernel Codebook Coding	68.75%	73.45%	76.90%	79.14%
Locality-constrained Linear Coding	66.14%	71.25%	79.21%	87.12%
Super Vector Coding	81.21%	86.27%	88.13%	90.90%
Improved Fisher Vector	85.00%	87.31%	91.25%	95.07%
RSDataset	**5**	**10**	**20**	**30**
Histogram Quantization Coding	61.20%	68.56%	77.50%	80.81%
Kernel Codebook Coding	70.56%	75.89%	79.31%	81.95%
Locality-constrained Linear Coding	68.66%	77.05%	85.60%	87.16%
Super Vector Coding	81.34%	83.89%	88.52%	91.73%
Improved Fisher Vector	83.51%	90.13%	92.97%	95.81%

5 to 30 for RSDataset. Note that the suggested number of training images in the literature for UCMereced it is 80, and for RSDataset is 30. As has been observed in multiple previous works [13, 15, 39], the performance can be improved considerably when the number of training samples is increased. On both datasets, and for all coding strategies a clear enhancement in classification accuracy is observed when the number of training images is increased.

Finally the complexity and speed of the encoding approaches are reviewed. Since BoVW framework consists of various steps, each step is evaluated separately:

1. *Key-point extraction and presentation* The 128-dimensional dense SIFT feature vector, has been adopted as the common descriptor for all the coding methods. The descriptors are computed using the ‘fast’ option in the publicly available VLFeat toolbox [35]. The fast version is not similar to the original SIFT descriptors and is slightly approximated, but it is from 30 to 70 times faster than SIFT. In our experiments computing ‘fast’ dense SIFT descriptors requires less than 0.5 seconds per image.
2. *Codebook generation and Encoding* To generate codebooks, an approximate nearest neighbor clustering algorithm [21] provided by the VLFeat toolbox is used. While histogram coding requires just the nearest neighbor to the descriptor, KCC and LLC need to search for more than one nearest neighbor per feature. The number of K nearest neighbors sought and the codebook size considerably increase the encoding time. In all our experiments we used $K = 5$ for both LLC and KCC, and the codebook size is fixed and is set to 16384. This results in an encoding time of around 0.5 seconds per image for histogram coding, 19 seconds for LLC and 24 seconds for KCC. Most of the encoding time is spent, however, on finding the K nearest neighbors, rather than the encoding. IFV on the other hand needs to be clustered using GMM. This does not cause a considerable overhead in clustering time. However due to their size, encoding IFV and SVC are quite slow for the codebook size of 16384. We thus used the suggested codebook size of 256 for IFV and 1024 for SVC in the literature. This configuration results in a 8-second encoding time for IFV and a 10-second encoding time for SVC. All the timings are for a 3.3GHZ Intel CPU where the implementations are done in C++/MATLAB.

5 Conclusion

The improved spatial resolution in HSR remote sensing imagery, provides more detailed information for land-use classification. Land-use scene categories, often cover multiple land-cover classes or ground objects. In addition, the higher spatial resolution may results in the increase of within-class spectral variation with the same surface features. As a consequence, pixel-base classification approaches can not fulfill this task anymore, and the approaches that represent the visual layout of land-use images flexibly are becoming popular for this purpose.

Local features, and BoVW in particular, is bridging this gap by providing an intermediate feature representation method. This paper investigated different configurations of BoVW framework for remote sensing land-use scene classification. The importance of this work lies in comparing different coding strategies for BoVW when its parameters are controlled. The detailed empirical evaluation of five coding schemes suggested that improved Fisher vector (IFV) [23] outperforms other coding strategies, though it introduces higher computational complexity. The performance of IFV is comparable to the state-of-the-art approaches [47, 48], where both local and global features are used in land-use scene classification. We also investigated the effect of codebook size and number of training images on UC Merced and RSDataset. Experimental findings showed that more training images and more visual words can improve the performance.

Acknowledgments This work is partially supported by a Discovery Grant to Professor Robert Bergevin from the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

1. Aksoy S, Koperski K, Tusk C, Marchisio G, Tilton JC (2005) Learning bayesian classifiers for scene classification with a visual grammar. IEEE Trans Geosci Remote Sens 43(3):581–589. doi:10.1109/TGRS.2004.839547
2. Avila S, Thome N, Cord M, Valle E, de A. Araújo A. (2013) Pooling in image representation: The visual codeword point of view. Comput Vis Image Underst 117(5):453–465. doi:10.1016/j.cviu.2012.09.007. http://www.sciencedirect.com/science/article/pii/S1077314212001737
3. Blaschke T (2010) Object based image analysis for remote sensing. ISPRS J Photogramm Remote Sens 65(1):2–16. doi:10.1016/j.isprsjprs.2009.06.004. http://www.sciencedirect.com/science/article/pii/S0924271609000884
4. Chen C, Zhang B, Su H, Li W, Wang L (2016) Land-use scene classification using multi-scale completed local binary patterns. SIViP 10(4):745–752. doi:10.1007/s11760-015-0804-2
5. Chen C, Zhou L, Guo J, Li W, Su H, Guo F (2015) Gabor-filtering-based completed local binary patterns for land-use scene classification. In: IEEE International conference on multimedia big data (bigMM), 2015, pp. 324–329. doi:10.1109/BigMM.2015.23
6. Chen S, Tian Y (2015) Pyramid of spatial relatons for scene-level land use classification. IEEE Trans Geosci Remote Sens 53(4):1947–1957. doi:10.1109/TGRS.2014.2351395
7. Csurka G, Dance CR, Fan L, Willamowski J, Bray C (2004) Visual categorization with bags of keypoints. In: In workshop on statistical learning in computer vision, ECCV, pp. 1–22
8. Dai D, Yang W (2011) Satellite image classification via two-layer sparse coding with biased image representation. IEEE Geosci Remote Sens Lett 8(1):173–176. doi:10.1109/LGRS.2010.2055033
9. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: IEEE Computer society conference on computer vision and pattern recognition, 2005. CVPR 2005. vol. 1, pp. 886–893 vol. 1. doi:10.1109/CVPR.2005.177
10. Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ (2008) LIBLINEAR: A library for large linear classification. J Mach Learn Res 9:1871–1874
11. Gao S, Tsang IWH, Chia LT (2010) Computer vision – ECCV 2010: 11th european conference on computer vision, heraklion, crete, Greece, september 5-11, 2010, proceedings, Part IV, chap. Kernel sparse representation for image classification and face recognition, pp. 1–14. Springer berlin heidelberg, berlin, heidelberg. doi:10.1007/978-3-642-15561-1_1
12. Hu J, Xia GS, Hu F, Sun H, Zhang L (2015) A comparative study of sampling analysis in scene classification of high-resolution remote sensing imagery. In: 2015 IEEE International geoscience and remote sensing symposium (IGARSS), pp. 2389–2392. doi:10.1109/IGARSS.2015.7326290
13. Huang Y, Wu Z, Wang L, Tan T (2014) Feature coding in image classification: a comprehensive study. IEEE Transactions on Pattern Analysis and Machine Intelligence 36(3):493–506. doi:10.1109/TPAMI.2013.113
14. Jaakkola TS, Haussler D (1999) Exploiting generative models in discriminative classifiers. In: Proceedings of the 1998 conference on advances in neural information processing systems II. MIT Press, Cambridge, MA, USA, pp 487–493. http://dl.acm.org/citation.cfm?id=340534.340715
15. Ken Chatfield Victor Lempitsky AV, Zisserman A (2011) The devil is in the details: an evaluation of recent feature encoding methods. In: Proceedings of the british machine vision conference, pp. 76.1–76.12. BMVA press. doi:10.5244/C.25.76
16. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: IEEE Computer society conference on computer vision and pattern recognition, 2006, vol. 2, pp. 2169–2178. doi:10.1109/CVPR.2006.68
17. Liu L, Wang L, Liu X (2011) In defense of soft-assignment coding. In: IEEE International conference on computer vision (ICCV), 2011, pp. 2486–2493. doi:10.1109/ICCV.2011.6126534
18. Lowe D (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110. doi:10.1023/B:VISI.0000029664.99615.94

19. Matas J, Chum O, Urban M, Pajdla T (2004) Robust wide-baseline stereo from maximally stable extremal region. Image Vis Comput 22(10):761–767. doi:10.1016/j.imavis.2004.02.006. British Machine Vision Computing 2002. http://www.sciencedirect.com/science/article/pii/S0262885604000435
20. McLachlan G, Peel D (2004) Finite mixture models. John Wiley & Sons
21. Muja M, Lowe DG (2009) Fast approximate nearest neighbors with automatic algorithm configuration. VISAPP (1) 2(331–340):2
22. Ojala T, Pietikäinen M, Harwood D (1996) A comparative study of texture measures with classification based on featured distributions. Pattern Recogn 29(1):51–59. doi:10.1016/0031-3203(95)00067-4. http://www.sciencedirect.com/science/article/pii/0031320395000674
23. Perronnin F, Liu Y, Sanchez J, Poirier H (2010) Large-scale image retrieval with compressed fisher vectors. In: IEEE Conference on computer vision and pattern recognition (CVPR), 2010, pp. 3384–3391. doi:10.1109/CVPR.2010.5540009
24. Philbin J, Chum O, Isard M, Sivic J, Zisserman A (2007) Object retrieval with large vocabularies and fast spatial matching. In: 2007 IEEE Conference on computer vision and pattern recognition, pp. 1–8. doi:10.1109/CVPR.2007.383172
25. Qi K, Wu H, Shen C, Gong J (2015) Land-use scene classification in high-resolution remote sensing images using improved correlatons. IEEE Geosci Remote Sens Lett 12(12):2403–2407. doi:10.1109/LGRS.2015.2478966
26. Sanchez J, Perronnin F (2011) High-dimensional signature compression for large-scale image classification. In: IEEE Conference on computer vision and pattern recognition (CVPR), 2011, pp. 1665–1672. doi:10.1109/CVPR.2011.5995504
27. dos Santos JA, Penatti OAB, da Silva Torres R, Gosselin PH, Philipp-Foliguet S, Falco A (2012) Improving texture description in remote sensing image multi-scale classification tasks by using visual words. In: 21St international conference on pattern recognition (ICPR), 2012, pp. 3090–3093
28. dos Santos JA, Penatti OAB, da Silva Torres R (2010) Evaluating the potential of texture and color descriptors for remote sensing image retrieval and classification. In: VISAPP (2), Pp. 203–208
29. dos Santos JA, da Silva Torres R (2013) Remote sensing image segmentation and representation through multiscale analysis. In: 26Th conference on graphics, patterns and images tutorials (SIBGRAPI-t), 2013, pp. 23–30. doi:10.1109/SIBGRAPI-T.2013.11
30. Shaw GA, Burke HHK (2003) Spectral imaging for remote sensing. Lincoln Laboratory Journal 14(1):3–28
31. Simonyan K, Parkhi OM, Vedaldi A, Zisserman A (2013) Fisher vector faces in the wild. In: British machine vision conference
32. Sivic J, Zisserman A (2003) Video google: a text retrieval approach to object matching in videos. In: Proceedings of the 9th IEEE international conference on computer vision, 2003, pp. 1470–1477 vol.2. doi:10.1109/ICCV.2003.1238663
33. Tuytelaars T (2010) Dense interest points. In: IEEE Conference on computer vision and pattern recognition (CVPR), 2010, pp. 2281–2288. doi:10.1109/CVPR.2010.5539911
34. Tuytelaars T, Van Gool L (2004) Matching widely separated views based on affine invariant regions. Int J Comput Vision 59(1):61–85. doi:10.1023/B:VISI.0000020671.28016.e8
35. Vedaldi A, Fulkerson B (2008) VLFEat: an open and portable library of computer vision algorithms. http://www.vlfeat.org/
36. Wang J, Yang J, Yu K, Lv F, Huang T, Gong Y (2010) Locality-constrained linear coding for image classification. In: IEEE Conference on computer vision and pattern recognition (CVPR), 2010, pp. 3360–3367. doi:10.1109/CVPR.2010.5540018
37. Xia GS, Yang W, Delon J, Gousseau Y, Sun H, Maître H. (2010) Structural High-resolution Satellite Image Indexing. In: Wagner B Székely W (ed) ISPRS TC VII Symposium - 100 years ISPRS, vol. XXXVIII. Vienna, Austria, pp 298–303. https://hal.archives-ouvertes.fr/hal-00458685
38. 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, pp. 1794–1801. doi:10.1109/CVPR.2009.5206757
39. Yang Y, Newsam S (2010) Bag-of-visual-words and spatial extensions for land-use classification. In: Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems, GIS ’10. ACM, New York, NY, USA, pp 270–279. doi:10.1145/1869790.1869829
40. Yu Q, Gong P, Clinton N, Biging G, Kelly M, Schirokauer D (2006) Object-based detailed vegetation classification with airborne high spatial resolution remote sensing imagery. Photogramm Eng Remote Sens 72(7):799–811

41. Zhang J, Cheng Z, Li T (2015) A bag-of-visual words approach based on optimal segmentation scale for high resolution remote sensing image classification. In: 2015 IEEE International geoscience and remote sensing symposium (IGARSS), pp. 1012–1015. doi:10.1109/IGARSS.2015.7325940
42. Zhang J, Li T, Lu X, Cheng Z (2016) Semantic classification of high-resolution remote-sensing images based on mid-level features. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 9(6):2343–2353. doi:10.1109/JSTARS.2016.2536943
43. Zhao L, Tang P, Huo L (2014) A 2-d wavelet decomposition-based bag-of-visual-words model for land-use scene classification. Int J Remote Sens 35(6):2296–2310. doi:10.1080/01431161.2014.890762
44. Zhao LJ, Tang P, Huo LZ (2014) Land-use scene classification using a concentric circle-structured multiscale bag-of-visual-words model. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 7(12):4620–4631. doi:10.1109JSTARS.2014.2339842
45. Zhao Y, Zhang L, Li P, Huang B (2007) Classification of high spatial resolution imagery using improved gaussian markov random-field-based texture features. IEEE Trans Geosci Remote Sens 45(5):1458–1468. doi:10.1109/TGRS.2007.892602
46. Zhou X, Yu K, Zhang T, Huang TS (2010) Computer vision – ECCV 2010: 11th european conference on computer vision, heraklion, crete, Greece, september 5-11, 2010, proceedings, Part V, chap. Image classification using super-vector coding of local image descriptors, pp. 141–154. Springer berlin heidelberg, berlin, heidelberg. doi:10.1007/978-3-642-15555-0_11
47. Zhu Q, Zhong Y, Zhao B, Xia GS, Zhang L (2016) Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery. IEEE Geosci Remote Sens Lett 13(6):747–751. doi:10.1109/LGRS.2015.2513443
48. Zou J, Li W, Chen C, Du Q (2016) Scene classification using local and global features with collaborative representation fusion. Inf Sci 348:209–226. doi:10.1016/j.ins.2016.02.021. http://www.sciencedirect.com/science/article/pii/S0020025516300755

Mana Shahriari has received her B.Eng. degree in Electrical Engineering from Shahid Bahonar University of Kerman, Iran and her M.Sc. degree in Artificial Intelligence from University of Southampton, UK in 2009 and 2011 respectively. She is currently enrolled in a PhD program at Laval University with the Computer Vision and Systems Laboratory. Her main research interests include image classification, content-based image retrieval and cognitive computer vision.

Robert Bergevin received the B.Eng. degree in Electrical Engineering and the M.A.Sc. degree in Biomedical Engineering from Polytechnique Montreal and the Ph.D. degree in Electrical Engineering from McGill University. His research interests are in image analysis and cognitive computer vision. His main works address generic modeling and recognition of objects in static images and tracking and modeling of people and animals in image sequences. Dr. Bergevin is a member of the Computer Vision and Systems Laboratory at Laval University, Quebec City, where he is a Professor in the Department of Electrical and Computer Engineering. He is a member of the Province of Quebec's Association of Professional Engineers (OIQ) and serves as Area Editor for the Computer Vision and Image Understanding journal.