Preprint - accepted for ICIAFS 2018 http://www.iciafs.org/iciafs/

Comparative Analysis of Different Features and Encoding Methods for Rice Image Classification

Vijayaratnam Emulse Nirmalan
College of Technology Jaffna
Sri Lanka
emulsen@gmail.com

Ruwan D. Nawarathna
Dept. of Statistics and Computer Science
Faculty of Science, University of Peradeniya
Sri Lanka
ruwand@pdn.ac.lk

Siyamalan Manivannan
Dept. of Computer Science
Faculty of Science, University of Jaffna
Sri Lanka
siyam@univ.jfn.ac.lk

***Abstract*—Rice is the most widely consumed staple food in Sri Lanka. In this paper, we present a comparative study of different features (SIFT, Multi-resolution Local Patterns, Local Color Histograms, and Random Projections) and feature encoding approaches (Bag-of-visual-words, Sparse Coding, Vector of Locally Aggregated Gradients, and Fisher Vectors) for classifying images containing rice grains. By analysing the performance of a classification model with two-fold cross validation on a dataset of 1000 images containing ten rice categories, we show that SIFT features with Fisher Vector encoding or with Vector of Locally Aggregated Gradients produces the best result (mean class accuracy of 97.9 ± 0.5). We found that increasing the size of the dictionary generally improves the classification performance for all the feature encoding approaches. The dataset we use is made public, and it can be accessed via http://www.csc.jfn.ac.lk/index.php/dataset/.**

***Index Terms*—image classification, features, feature encoding, bag-of-visual-words, SIFT and Fisher Vectors**

I. Introduction

Rice is the most widely consumed staple food in Sri Lanka. There are many varieties of rice exist in the market. They differ from each other mainly based on the features such as size, shape and color. Automatic identification of rice varieties would be very useful for the consumers, for example, a consumer can take a photo of the rice displayed in the supermarket using his/her mobile device and can get more details of it, such as its online price as he/she can automatically identify its category.

In the last decade, various approaches have been proposed for rice image classification, which use different features and classifiers, for example, Kaur et. al. [1] used shape-based features and Support Vector Machine (SVM) classifier, Mousavirad el. al. [2] used morphological features and Neural Networks (NN) classifier, Chaugule et. al. [3] used a set of texture and shape features and a NN classifier, Sumaryanti et. al. [4] used color, texture and morphological features with NN classifier. Most of these approaches try to categorize individual rice grains instead of classifying bulk of rice. Also, based on the literature it is difficult to identify which features and classifiers are best suited for rice image classification as different methods are tested on different small-scale datasets. In addition, to best of our knowledge, feature encoding approaches such as Fisher Vectors [5] have not been applied for rice image classification, although they have been widely applied in other domains [6]. Therefore, in this paper we compare different features and feature encoding methods for rice image classification. In contrast to most of the existing work, we focus on categorizing bulk of rice instead of individual rice grains. Note that in the work related to classifying individual rice grains the accuracy of the system heavily depends on how well individual rice grains are segmented. Since we focus on bulk of rice our system avoids segmenting individual grains from images which contain rice, hence, higher accuracy can be obtained. As an additional contribution we introduced a new dataset with 1,000 rice images from 10 different categories and make it publicly available, which enable other researchers to apply their techniques on this dataset and compare with our technique easily. A few sample images from the dataset are shown in Figure 1.

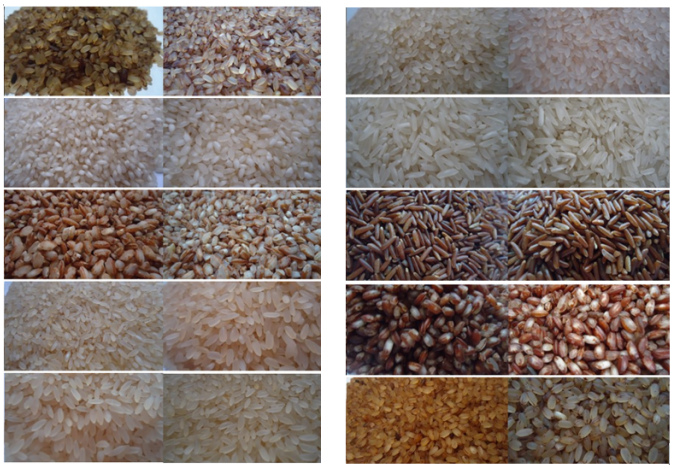


Fig. 1. Example images from our dataset: Each row in the two columns show images from different rice categories.

II. Related Work

As mentioned in Section I most of the existing work (e.g. [1]–[4], [7]–[9]) focus on classifying individual rice grains into one of the predefined categories based on the features such as shape, texture and color and classifiers such as NN and SVM. In these approaches, first, individual rice grains are segmented

using image processing techniques and then features such as color, shape, and texture were extracted from the segmented images. Finally the extracted features were used to train a classifier. These approaches assume that each image contains one or only few grains so that segmenting them would be easy, or they assume that the images are taken under controlled environments, e.g., individual rice grains were photographed on a white background [7]. However in reality rice are in bulk quantity and their background can vary, and the images can be taken at different scales and illumination conditions (see Figure 1), hence segmenting individual grains is a difficult task.

In the literature of rice image classification, usually, simple features which capture the basic properties of individual rice grains were used, for example, morphological features such as area, perimeter, major and minor axis length of segmented rice grains were used in [2]. On the other hand, features such as SIFT, Random Projections [10], etc., have been widely used in computer vision together with the feature encoding approaches such as Bag-of-words and Fisher Vectors. These features and encoding methods perform well for various problems such as image classification [11], and segmentation [12]. However, to best of our knowledge these features or feature encoding approaches have not yet explored for rice image classification. Therefore, in this work we focus on the applicability of these features and encoding approaches for automatic rice image classification.

III. Methodology

An overview of the proposed system for generating feature representation for a given rice image is illustrated in Figure 2. Firstly, each image was pre-processed. Local features were then extracted and a feature encoding method (e.g. bag-of-visual-words) was employed to aggregate the local features into an image representation. A support vector machine was then used to classify the rice images. The following sections describe these steps in detail.

A. Image pre-processing

Since the rice images were taken under different illumination conditions a pre-processing step is necessary to normalize the images. An intensity normalization method was used, where the intensity values in each image were linearly rescaled so that 2% of pixels in each image became saturated at low and high intensities [6].

Since the original images are in high resolution (1080x1920 pixels), processing them will take a significant amount of time. To make the processing faster each image was resized by preserving its height to width aspect ratio such that its maximum dimension (height or width) becomes 400 pixels.

B. Local feature extraction

In image classification, image descriptors/features play an important role as they capture image/region properties, such as color, shape, edges, texture, etc. In general, there are two approaches to describe an image using descriptors, *global* and *local*. The global descriptor captures the overall statistics of an image, e.g. color histogram computed from an image contains rice grains. However, since this is a global representation, it may fail to capture the local properties of the image, such as the shape, color or texture properties of individual rice grains. In rice image classification local image properties such as local shape, texture, etc., are more important than the global ones. Local descriptors (e.g., SIFT) can be used to capture the local image properties, and they are designed to be robust to image transformations.

Mainly there are two sampling methods generally used for local feature extraction (i) *dense sampling*, where the feature extraction is based on a regular grid of points placed over the images, and (ii) *interest points*, where special points in the images are identified by a detector (e.g., Harris detector [13]) and feature descriptors computed around those points. Dense feature sampling seems to lead to better performance compared to interest point detectors for image classification [14]. Therefore in this work features were extracted using dense sampling.

Various local descriptors have been proposed in the literature to effectively capture the local image properties, for example, SIFT descriptors capture local shape/texture, LBP [15] descriptors capture texture features. In this research we compare the classification performance of four different descriptors, namely, SIFT, Multi-resolution Local Patterns, Local Color Histograms and Random Projections.

To capture information at different scales from each image these features were extracted from two different sizes of patches, 16x16 and 32x32 pixels with a step size of 4 pixels. Since rice images are in color, for each feature type we compute the following features separately from each color channel of the RGB patch and concatenate them to get a feature representation for that patch. The following sections describe these features in detail.

1) SIFT: It captures the histogram representation of local image derivatives inside small image regions. SIFT has been widely used in computer vision for image classification. The size of the SIFT feature to represent a RGB color patch is $128 \times 3 = 384$.

2) Multi-resolution Local Patterns (mLP): This descriptor was proposed by Manivannan et. al. in [6]. It is a non-binarized, multi-resolution version of the well-known Local Binary Patterns (LBP) descriptor. Here we use a 3-resolution version of the mLP descriptor, with 8, 12, and 16 sampling points in the first, second and the third resolutions respectively. This leads to a feature dimension of $36 \times 3 = 108$ for each RGB patch.

3) Random projection (RP): It is a dimensionality reduction technique, successfully used as a texture descriptor in [10] for texture image classification. It projects patch intensity vectors from the original patch-vector space to a compressed space using randomly chosen projection vectors. In this work, we project each of the linearized RGB patch vectors to a dimension of 200. For example, when 16×16 patch is

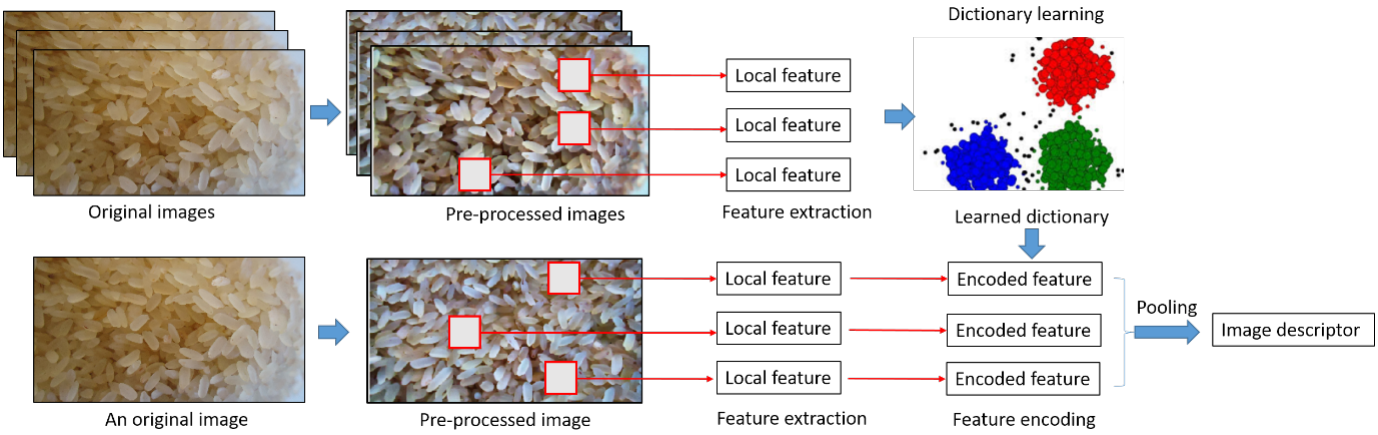


Fig. 2. An overview of the system for generating the image-level feature representation: dictionary learning from training images (first row) and feature encoding to obtain the image-level feature representation (second row).

considered, we project the linearized RGB patch of dimension $768 (= 16 \times 16 \times 3)$ to a compressed space of dimension 200.

4) Local Color Histograms (LCH): From each color channel of the RGB color space of each local patch, we compute an intensity histogram with 256 bins. Histograms computed from each color channel were then concatenated to get a feature vector of size $256 \times 3 = 768$ to represent each patch.

Values of each parameter corresponding to these four feature descriptors were decided experimentally.

C. Feature Encoding

Feature encoding is a way to compute image representations by aggregating the local features extracted from each image. We compare the following four feature encoding methods: bag-of-words (BoW), Sparse Coding (SC), Fisher vectors (FV), and Vectors of Locally Aggregated Descriptors (VLAD). In all of these methods, first, a dictionary was built from the local features extracted from the training images. This dictionary was then used to encode the local features extracted from each image to get image representations. Dictionary elements are often referred to as *clusters* or *visual words*. Each of these methods is described briefly as follows.

1) Bag-of-words (BOW): It has been widely applied for image classification [11]. In BOW image representation is computed as a frequency histogram, where i^{th} bin of this histogram represents the number of local features which are assigned to i^{th} visual word.

2) Sparse coding (SC): In SC [11] each local image descriptor is reconstructed using a weighted combination of a few dictionary elements. SC has shown improved performance over BOW for image classification [11]. In this work, we use an efficient variant of SC called the Locality-constrained linear coding (LLC) [16], which enforces locality instead of sparsity. LLC utilizes the local linear property of manifolds to project each descriptor into its local coordinate system.

To aggregate the local features two kinds of pooling, *max* and *sum*, were used in the literature for SC [16]. Therefore, we used two variants of SC, which are *SC-sum* which uses sum pooling and *SC-max* which uses max-pooling, respectively.

3) Fisher Vector (FV): FV [5] has shown improved performance over BOW and SC for image classification. Compared to BOW, FV captures additional information about the distribution of the local features inside each cluster. In FV the dictionary is built using a Gaussian Mixture Model (GMM). Each cluster is then represented based on the derivative of the GMM with respect to its parameters. In BOW and SC, the size of the image representation is the same as the number of dictionary elements $(= K)$. However, in FV the size of the image representation is much higher, i.e., $2KD$, where D represents the size of the local feature used, e.g., for SIFT, $D = 384$.

4) Vectors of Locally Aggregated Descriptors (VLAD): This encoding method was proposed in [17] as a simple approximation to FV. In VLAD the dictionary is learned using the k-means clustering algorithm. Each image in VLAD is represented by a vector of size KD.

D. Normalization of Image Representations

The number of local features extracted from each image vary because of different sizes of the images. Therefore a normalization step is necessary to make the image representations comparable. In this work, we applied the *L2-and-power normalization* proposed in [5] for normalizing the computed image representations by the above feature encoding methods.

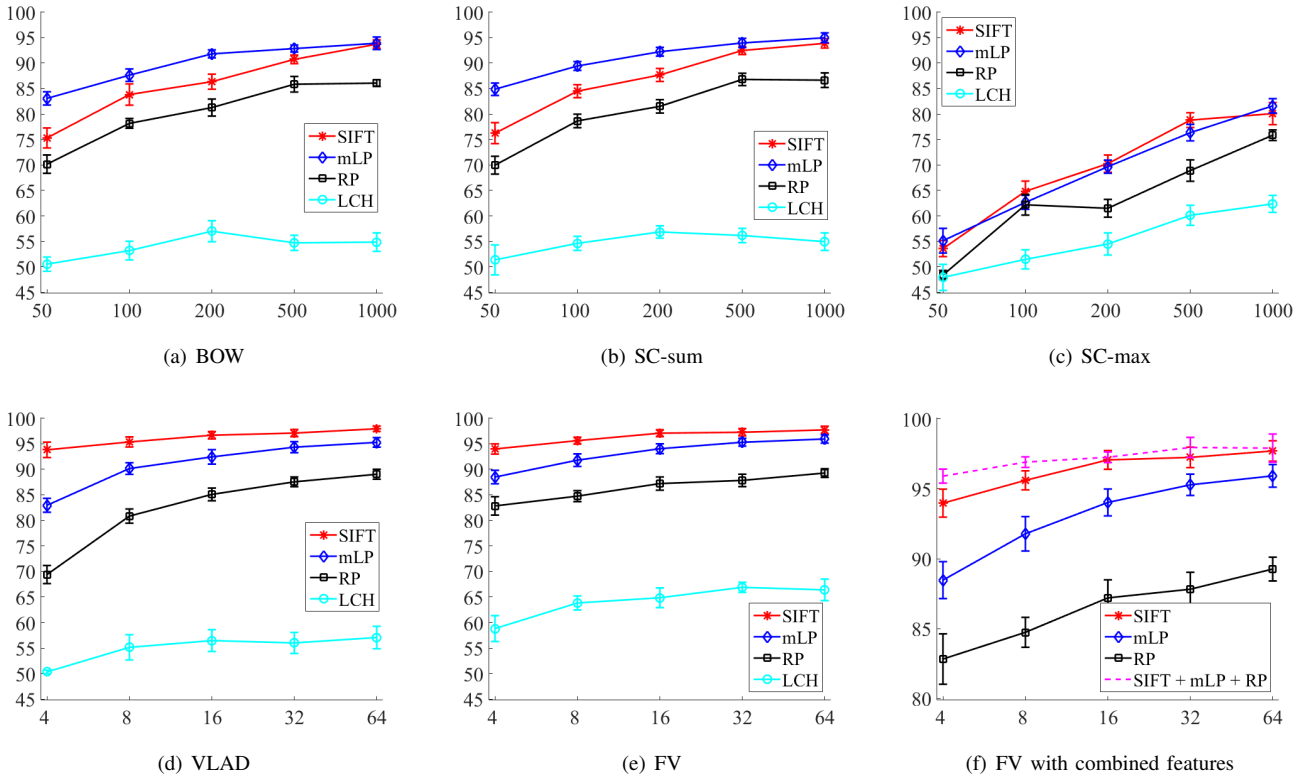
Let $\mathbf{z}_i \in \mathbb{R}^d$ represents the image-level representation of an image I_i, where d is the size of the representation, the *L2-and-power* normalizations can be given as.

$$\mathbf{z}_i \leftarrow \frac{sign(\mathbf{z}_i)|\mathbf{z}_i|^{\frac{1}{2}}}{\|\mathbf{z}_i\|_2} \tag{1}$$

where $|\mathbf{z}_i|^{\frac{1}{2}}$ applies the square root to each component of $\mathbf{z}_i$.

E. Classification

A one-vs-rest, multi-class linear SVM was used as the classifier. We used the `liblinear` [18] implementation of



(a) BOW (b) SC-sum (c) SC-max

(d) VLAD (e) FV (f) FV with combined features

Fig. 3. Performance of different features and encoding methods for different sizes of the dictionaries. Each graph plots Mean Class Accuracy (MCA) (vertical axis) vs size of the dictionary (horizontal axis) for each encoding method, (a) Bag of words (BOW), (b) Sparse coding (SC) with sum-pooling, (c) SC with max-pooling, (d) Vectors of locally aggregated descriptors (VLAD), (e) Fisher vector (FV), and (f) FV with SIFT-RP-mLP combined. (a)-(e) show the performance of SIFT, Multi-resolution local patterns (mLP), Random projection (RP) and Local color histograms(LCH) whereas (f) shows the performance of SIFT, mLP, RP, and SIFT-RP-mLP combined.

the SVM classifier for this purpose. The cost parameter of the SVM classifier was determined based on applying a two-fold cross-validation on the training set.

IV. Experiments

A. Rice Images Dataset

Ten verities of rice were obtained from the local market in Sri Lanka. These rice were imaged under different illumination conditions, view point changes and under different scales. Hundred images were taken from each category, this leads to a total of 1000 ($= 10 \times 100$) images. Some of the images from our dataset are shown in Figure 1.

B. Experimental settings

The public library, `vlfeat` [19], was used for SIFT feature extraction, dictionary learning (k-means, GMM) and feature encoding. For SC, we used the implementation of LLC from [16]. K-means with 200,000 randomly sampled instances of each type of local feature was used to build the dictionaries for BOW, SC and VLAD methods.

Mean Class Accuracy (MCA) was used as the evaluation metric, and can be given as

$$MCA = \frac{1}{C}\sum_{c=1}^{C} CCR_c \qquad (2)$$

where C is the number of classes ($C = 10$), CCR_c is the correct classification rate for class c.

We applied a two-fold cross validation (repeated 5 times) and report the mean and the standard deviations of the MCA obtained over these iterations.

C. Results and discussions

Results of different combination of features and encoding methods for different sizes of dictionaries are shown in Figure 3 and Table I. For all the features (except LCH) and encoding methods, MCA value improves as the dictionary size increases. Regardless of the feature encoding method used, SIFT and mLP features perform better than others. For SC, sum pooling (SC-sum) performs much better than max pooling (SC-max). BOW and SC-sum give similar performance. VLAD and FV perform better than BOW and SC even with smaller dictionaries. The sizes of the image representations are given in Table I. SIFT feature with FV encoding gave the best performance compared to other features when the dictionary size is set to 64. LCH gave worst performance suggesting that color is not a discriminative feature for rice image classification.

Figure 3(f) shows the classification performance when the features (SIFT, mLP, and RP) were combined with FV encoding. This feature combination gives better performance than all

TABLE I

PERFORMANCE (MEAN $\pm$ STD.) OF DIFFERENT FEATURES WITH DIFFERENT ENCODING METHODS: THE DICTIONARY SIZE IS SET TO $1,000$ FOR BOW, SC WITH SUM AND MAX POOLING, AND THE DICTIONARY SIZE IS SET TO 64 FOR THE VLAD AND FV APPROACHES. THE SIZE OF IMAGE REPRESENTATIONS FOR DIFFERENT FEATURES WITH DIFFERENT ENCODING APPROACHES ARE GIVEN INSIDE PARENTHESIS.

Type of the Feature	Type of Encoding				
	BOW	SC-sum	SC-max	VLAD	FV
SIFT	$93.70 \pm 0.86\,(1,000)$	$93.86 \pm 0.92\,(1,000)$	$80.08 \pm 2.17\,(1,000)$	$\mathbf{97.92 \pm 0.52}\,(24,576)$	$\mathbf{97.70 \pm 0.73}\,(49,152)$
mLP	$93.88 \pm 1.23\,(1,000)$	$94.96 \pm 0.96\,(1,000)$	$81.60 \pm 1.41\,(1,000)$	$95.26 \pm 0.96\,(6,912)$	$95.92 \pm 0.81\,(13,824)$
RP	$86.06 \pm 0.66\,(1,000)$	$86.64 \pm 1.43\,(1,000)$	$75.84 \pm 1.04\,(1,000)$	$89.02 \pm 0.97\,(12,800)$	$89.26 \pm 0.85\,(25,600)$
LCH	$54.86 \pm 1.81\,(1,000)$	$54.96 \pm 1.72\,(1,000)$	$62.36 \pm 1.67\,(1,000)$	$57.08 \pm 2.20\,(49,152)$	$66.40 \pm 2.10\,(98,304)$

individual features when the dictionary size is small ($K = 4$). However, when the dictionary size is large ($K > 8$) this combination gives similar performance to the best performing feature (i.e., SIFT). This is mainly due to the saturation of the classification performance.

V. CONCLUSIONS

In this paper, we compared different features and feature encoding methods in Computer Vision for rice image classification. It shows that SIFT feature with Fisher Vector (FV) encoding and SIFT feature with Vector of Locally Aggregated Descriptors (VLAD) performs better than other features and feature encoding methods. A state-of-the-art accuracy of 97.92 ± 0.52 was obtained using a one-vs-rest, multi-class linear SVM.

REFERENCES

[1] H. Kaur and B. Singh, "Classification and grading rice using multi-class SVM," *International Journal of Scientific and Research Publications*, vol. 3, no. 4, 2013.

[2] S. MousaviRad, F. A. Tab, and K. Mollazade, "Design of an expert system for rice kernel identification using optimal morphological features and back propagation neural network," *International Journal of Applied Information Systems*, vol. 3, no. 2, 2012.

[3] A. Chaugule and S. N. Mali, "Evaluation of texture and shape features for classification of four paddy varieties," *Journal of Engineering*, 2014.

[4] L. Sumaryanti, A. Musdholifah, and S. Hartati, "Digital image based identification of rice variety using image processing and neural network," *TELKOMNIKA Indonesian Journal of Electrical Engineering*, vol. 16, no. 1, pp. 182–190, 2015.

[5] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," in *European Conference on Computer Vision*, 2010.

[6] S. Manivannan, W. Li, S. Akbar, R. Wang, J. Zhang, and S. J. McKenna, "An automated pattern recognition system for classifying indirect immunofluorescence images of hep-2 cells and specimens," *Pattern Recognition*, vol. 51, pp. 12–26, March 2016.

[7] H. S. Gujjar and D. M. Siddappa, "A method for identification of basmati rice grain of india and its quality using pattern classification," *International Journal of Engineering Research and Applications*, vol. 3, no. 1, pp. 268–273, 2013.

[8] M. R. Siddagangappa and A. H. Kulkarni, "Classification and quality analysis of food grains," *IOSR Journal of Computer Engineering*, vol. 16, no. 4, pp. 1–10, 2014.

[9] A. R. Pazoki, F. Farokhi, and Z. Pazoki, "Classification of rice grain varieties using two artificial neural networks (MLP and neuro-fuzzy)," *The journal of Animcal and Plant Sciences*, vol. 24, no. 1, pp. 336–343, 2014.

[10] E. Bingham and H. Mannila, "Random projection in dimensionality reduction: applications to image and text data," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2001, pp. 245–250.

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

[12] W. Li, S. Manivannan, J. Zhang, E. Trucco, and S. J. McKenna, "Gland segmentation in colon histology images using hand-crafted features and convolutional neural networks," in *International Symposium on Biomedical Imaging (ISBI)*, 2016.

[13] C. Harris and M. Stephens, "A combined corner and edge detector," in *In Proceedings of Fourth Alvey Vision Conference*, 1988.

[14] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," in *European Conference on Computer Vision*, 2006.

[15] T. Ojala, M. Pietikainen, and D. Harwood, "Performance evaluation of texture measures with classification based on Kullback discrimination of distributions," in *International Conference on Pattern Recognition*, 1994.

[16] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *IEEE Computer Vision and Pattern Recognition*, 2010.

[17] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *IEEE Computer Vision and Pattern Recognition*, 2010.

[18] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[19] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," http://www.vlfeat.org/, 2008.