8-2016

A novel digital image classification algorithm via low-rank sparse bag-of-features model

Xiu-Ming ZOU

Huai-Jiang SUN

Sai YANG

Yan ZHU
Singapore Management University, yanzhu@smu.edu.sg

A novel digital image classification algorithm via low-rank sparse bag-of-features model

Zou Xiu-Ming[1,2*], Sun Huai-Jiang[1] , Yang Sai[3], Zhu Yan[4]
[1] School of Computer Science and Engineering, Nanjing University of Science and technology, Nanjing, 210094, China
[2] School of Computer Science and Technology, Huaiyin Normal University, Huaian, 223300, China
[3] School of Electrical Engineering, Nantong University, Nantong, 226019, China
[4] Integrated Information Technology Services, Singapore Management University, Singapore, 188065, Singapore
* Corresponding Author, email:brightzou@126.com

Journal of Digital Information Management

ABSTRACT: *Bag-of-features (BoF) is one of the most well-known methods used to represent digital image features because of its simplicity and efficiency. A variety of improved algorithms have been employed to enhance the performance of BoF in characterization. However, challenges in the application of BoF in the field still exist. This study focused on BoF by decomposing local features and presented a novel framework for BoF on the basis of low-rank and sparse matrix decomposition to obtain a more robust and discriminative digital image classification. First, the local feature matrix of a digital image is decomposed into a low-rank matrix and a sparse matrix. Then, the BoF model was constructed in each part. Finally, the multiple kernel learning method was applied to combine the two models and the digital images were classified by using the support vector machine. Compared with existing methods in five public data sets, results show that the method proposed in this study is superior to the baseline algorithm and other coding algorithms by improving local features, with an improved classification performance of 17.68% in maximum and 0.01% in minimum. Compared with similar methods (such as leveraging the low-rank and sparse matrix decomposition and group sparse coding for image classification), this method is superior, with an improved classification performance of 2.76% in maximum and 0.08% in minimum, and obtains the highest average correct rate of classification. Therefore, the proposed method in this study is effective in improving the BoF in the feature extraction stage and has a better image classification performance.*

Subject Categories and Descriptors
H.2.8 [Database Applications]: Image database; **I.3.5 [Image Processing and Computer Vision]:** Image Display; **I.4.10 [Image Representation]**

General Terms
Classification algorithm; Features; Image Processing

Keywords: low-rank and sparse matrix decomposition, bag-of-features, multiple kernel learning, digital image classification

Received: 1 May 2016, Revised: 27 June 2016, Accepted: 3 July 2016

1. Introduction

Automatic classification based on the contents of digital images has become an important aspect in the field of digital image processing to organize and manage massive amounts of digital image information rationally and efficiently. At the same time, digital image classification has recently become one of the hot topics in the research

on computer vision. The basic process of digital image classification involves establishing descriptions of digital image contents, learning digital image types by machine learning, and classifying unknown digital images with the acquired models. Among all of the methods representing digital image contents, the bag-of-features (BoF) model has recently become a well-known digital image representation method by expressing vector sets indirectly with a visual dictionary, expressing a digital image as histograms of visual word frequencies, and narrowing down the "semantic gap" between bottom visual features and top semantic features.

The core idea of the BoF model is to encode the local descriptive vectors of a digital image with well-trained off-line visual dictionaries and use the statistical values of the coded vectors to represent the features of the digital image. The current BoF models usually construct the middle semantic features of a digital image by directly using the local features. However, the low-rank and sparse parts of the local features represent the pertinent and unique properties of a digital image, respectively. Better identification of classification performance with more robustness can be achieved by decomposing the low-rank and sparse features. For this purpose, this study proposes a novel expression form of the BoF model on the basis of low-rank and sparse decomposition.

2. State of The Art

Currently, research on BoF mainly focuses on coding and pooling. Originally, this model is a hard coding method based on voting strategy, which is simple but highly sensitive to reconstruction errors. A more robust voting strategy is soft coding[1], which assigns local features to visual words. The sparse coding method proposed by Yang and Yu[2] can equally reduce reconstruction errors; however, the method is time-consuming and fails to ensure coding consistency. As a result, in Reference [3], the Laplacian matrix was added to the objective function to improve consistency in sparse coding. In References [4,5,6], contiguous visual words were used to represent local features to maintain coding sparsity and improve coding speed. Except for the research presented in Reference [3], the aforementioned coding methods separately code local features, ignoring the spatial contextual information among them. Therefore, Shabou et al.[7] proposed a locality-constrained and spatially regularized coding method. Other methods include the low-rank sparse coding by Zhang and Ghanem[8] and the group-salient coding by Wu et al.[9]. Moreover, BoF originally represents visual words as histograms at the pooling stage, which significantly restricts the representation features of the model. As a result, Lazebnik et al.[10] proposed spatial pyramid matching, which hierarchically partitions a digital image into several subsections, clusters coding vectors in each subsection, and concatenates the subsections into a vector to represent the features in each hierarchy of the digital image. This method has become a widely used representation

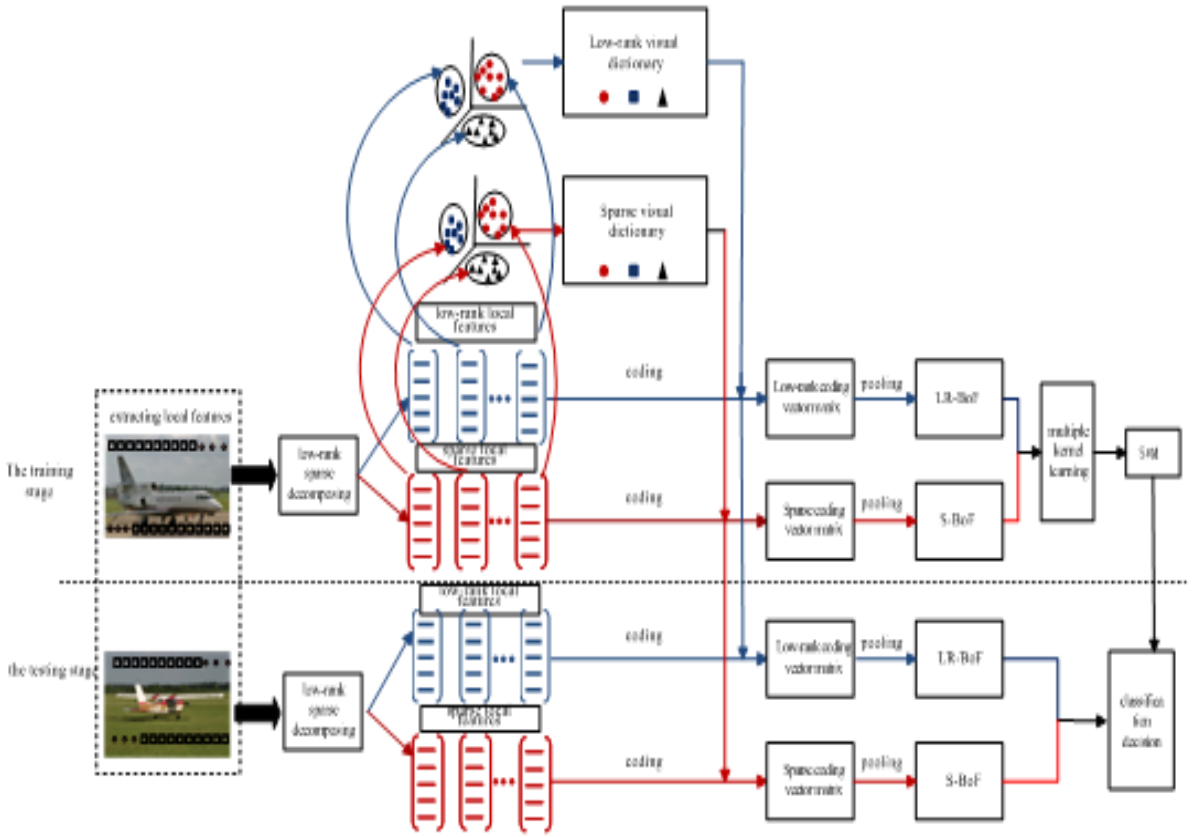

Figure 1. General framework for digital image classification

framework for digital image features because of its superior results in actual application.

With the rank of the matrix as a measurement of sparsity, low-rank matrix recovery has become a new high-dimensional data analysis tool and has been applied in video monitoring, face recognition, digital image shadow, and illumination normalization[11]. Zhang and Liu[12] decomposed the global features of an image into low-rank and sparse parts, combined the base vectors of the dictionary constructed from the two parts, and encoded and classified the original global features of the image. Zhang and Ma[13] extended the model by extracting the matrices of local features from an image and enhanced its performance in the first procedure of the BoF model. Combining the visual dictionaries constructed from the channels of low-rank and sparse features, this method establishes BoF models of the original local features on the basis of the dictionaries. Compared with the BoF models directly constructed from local features, this method has a more robust and differentiating classification performance because the two parts represent the pertinent and unique properties of the image. In contrast to the method presented in Reference [13], this study not only constructs visual dictionaries from the low-rank and sparse channels but also encodes the low-rank and sparse features by using the corresponding dictionaries, establishes low-rank bag-of-features (LR-BoF) and sparse bag-of-features (S-BoF) from the two channels, fuses the two BoF models and classifies the image on the basis of the support vector machine (SVM), which is a multiple kernel learning method.

The remainder of this paper is organized as follows: In Section 3, we establish the overall framework for the digital image classification method and describe the proposed LR-BoF model in detail. In Section 4, we conduct experimental simulation by applying the proposed model in the five public data sets and compare our results with those of other methods. Section 5 concludes this paper. We summarize our main results and provide some prospects for future work.

3. Methodology

The digital image classification method is shown in Figure.1. The specifics are as follows: In the digital image data set $D = \{d_1, d_2, \ldots, d_N\}$, each image is described as a matrix composed of many local features after local feature extraction. For convenience sake, the local feature matrix of each picture is marked as $X \in R^{D \times T}$, where D denotes the dimensions of local features and T denotes the number of local features in the image. For the essential low-rank data recovered from the observed data from greater sparse noise pollution, the optimized model can be described by the following formula:

$$\min_{Z,E} rank(\boldsymbol{Z}) + \gamma \left\| \boldsymbol{E} \right\|_0 \quad s.t. \quad \boldsymbol{X} = \boldsymbol{Z} + \boldsymbol{E} \tag{1}$$

In Formula (1), $\boldsymbol{Z} \in R^{D \times T}$ corresponds to the low-rank matrix part, whereas $\boldsymbol{E} \in R^{D \times T}$ corresponds to the part with sparse features. For the high non-convex optimization in the previously presented problematic model, by convex relaxation[14], that is, substituting Norm l_1 for Norm l_0, of the nuclear norm for rank function, a convex optimization model can be obtained as follows:

$$\min_{Z,E} \left\| \boldsymbol{Z} \right\|_* + \gamma \left\| \boldsymbol{E} \right\|_1 \quad s.t. \quad \boldsymbol{X} = \boldsymbol{Z} + \boldsymbol{E} \tag{2}$$

In Formula (2), $\|\cdot\|$ stands for Norm l_1, which expresses the sum of the absolute values for all of the elements in the matrix, and $\|.\|_*$ stands for the nuclear norm, which expresses the sum of singular values of the matrix. Optimization can be achieved by using the augmented Lagrange multiplier algorithm[15].

The sets of low-rank and sparse features of all of the training images are expressed as $\boldsymbol{Z}_{train} = \{\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_M\} \in R^{D \times M}$ and $\boldsymbol{E}_{train} = \{\boldsymbol{E}_1, \boldsymbol{E}_2, \ldots, \boldsymbol{E}_M\} \in R^{D \times M}$, respectively. One subset is randomly selected from $\boldsymbol{Z}_{train}$ and $\boldsymbol{E}_{train}$. The low-rank visual dictionary $\boldsymbol{V}_Z = [\boldsymbol{v}_{Z1}, \boldsymbol{v}_{Z2}, \ldots, \boldsymbol{v}_{ZK}]^T \in R^{D \times K}$ and the sparse visual dictionary $\boldsymbol{V}_E = [\boldsymbol{V}_{E1}, \boldsymbol{V}_{E2}, \ldots, \boldsymbol{V}_{EK}]^T \in R^{D \times K}$ are obtained after *K*-means clustering of the subset, where *D* stands for the dimensions of low-rank and sparse features, which are the same as the dimensions of the local features, whereas *K* stands for the number of visual words. Closely linking feature extraction[16,17] and feature pooling[18], coding is the core procedure in the BoF model. Among the current classical coding algorithms, locality-constrained linear coding (LLC)[4] has been widely used because of its speed and effectiveness. For convenience sake, assuming that the set of the low-rank features of one image in the data set is expressed as $\boldsymbol{Z} = [\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_T] \in R^{D \times T}$ and the set of sparse features thereof is expressed as $\boldsymbol{E} = [\boldsymbol{E}_1, \boldsymbol{E}_2, \ldots, \boldsymbol{E}_T] \in R^{D \times T}$, where $\boldsymbol{Z}_t$ and $\boldsymbol{E}_t$ stand for the *t* th low-rank and sparse features, respectively, the coding of $\boldsymbol{Z}_t$ by LLC is expressed as follows:

respectively, the coding of $\boldsymbol{Z}_t$ by LLC is expressed as follows:

$$\boldsymbol{u}_{zt} = \min_{u_{Zt} \in R^{1 \times K}} \left\| \boldsymbol{Z}_t - \boldsymbol{V}_Z \boldsymbol{u}_{zt} \right\|^2 + \lambda \left\| \boldsymbol{d}_{Zt} \Theta \boldsymbol{u}_{zt} \right\|^2 \quad s.t. \quad \left| \boldsymbol{u}_{zt} \right|_1 = 1 \tag{3}$$

The coding of $\boldsymbol{E}_t$ by LLC is expressed as follows:

$$\boldsymbol{u}_{Et} = \min_{u_{St} \in R^{1 \times K}} \left\| \boldsymbol{E}_t - \boldsymbol{V}_E \boldsymbol{u}_{St} \right\|^2 + \lambda \left\| \boldsymbol{d}_{Zt} \Theta \boldsymbol{u}_{Et} \right\|^2 \quad s.t. \quad \left| \boldsymbol{u}_{St} \right|_1 = 1 \tag{4}$$

In Formulas (3) and (4), $\boldsymbol{u}_{Zt}$ and $\boldsymbol{u}_{Et}$ stand for the corresponding coding vector of $\boldsymbol{Z}_t$ and $\boldsymbol{E}_t$, respectively;

Θ stands for the distance vector, $\boldsymbol{d}_{Zt}$,$\boldsymbol{d}_{Et}$ stand for the products of the corresponding elements of $\boldsymbol{u}_{Zt}$,$\boldsymbol{u}_{Et}$; and $\boldsymbol{d}_{Zt}$,$\boldsymbol{d}_{Et}$ stand for the distance vectors between $\boldsymbol{Z}_t$,$\boldsymbol{E}_t$ and the vectors of visual words, expressed as follows:

$$d_{zt} = \exp\left(\frac{dsit\,(Z_t, V_z)}{\sigma}\right)$$
$$d_{Et} = \exp\left(\frac{dist\,(E_t, V_s)}{\sigma}\right) \quad (5)$$

In Formula (5), σ stands for the weight of the fall speed of the locality-constrained factors. Generally, the fast approximation algorithm is used to optimize Formula (5) and accelerates encoding to represent each local feature by selecting *k* vectors of contiguous visual words.

At the pooling stage, the spatial pyramid model is used to partition an image into three levels, namely, L_0, L_1, and L_2, with the number of subsections at each level of the image being 1 × 1, 4 × 4, and 16 × 16, and to pool the coding vectors in subsections at each level maximally. Assuming that T_i coding features exist in the ***i*** subsection at level *l* ($l = L_0, L_1, L_2$), the pooled low-rank and pooled sparse features at the section can be expressed as follows:

$$\mathbf{B_{Zli}} = \max_{t=1,2,...T_i} \mathbf{u_{Zt}}$$
$$\mathbf{B_{Sli}} = \max_{t=1,2,...T_i} \mathbf{u_{St}} \quad (6)$$

In Formula (6), $\boldsymbol{B}_{Zli}$ and $\boldsymbol{B}_{Sli}$ stand for the *K*-dimension feature vectors. The low-rank feature $\boldsymbol{B}_{Zl}$ and sparse feature $\boldsymbol{B}_{El}$ at level *l* of the image can be obtained by linking all of the features at this level. LR-BoF and S-BoF are obtained by linking the features at the three levels.

Assuming that the kernel function of feature $\boldsymbol{B}_Z$ of LR-BoF and that of feature $\boldsymbol{B}_E$ of S-BoF are K_Z and K_E, respectively, and the corresponding weights are C_Z and C_E, respectively, the similarities between any two images can be expressed as follows:

$$K(i,j) = C_Z K(B_{Zi}, B_{Zj}) + C_E K(B_{Ei}, B_{Ej}) \quad (7)$$

If the SVM[19,20] is selected for classification differentiation of the image, then, in the second classification, the optimized objective function can be written as follows:

$$\min \frac{1}{2}\left(\frac{1}{C_z}\left|\left| wz \right|\right|^2 + \frac{1}{C_E}\left|\left| W_E \right|\right|^2 C\sum_{i-1}^{n} \xi\right.$$
$$s.t. y_i(w^T_Z \phi(B_{zi}) + w^T_E \phi(B_{Ei}) + b) \geq 1 - \xi_i. \quad (8)$$
$$\xi_i \geq 0, \quad i=1,2,....N$$
$$C_z, C_E \geq 0 \;\; C_z + C_E = 1$$

In Formula (8), y_i, y_j are the corresponding type labels of the images, *C* is the penalty coefficient, ξ_i is the relaxation variable, *w* is the parameter of the classifier, and *N* is the number of image samples. By solving the objective function in Formula (8) with the block coordination descent algorithm, we can arrive at the result that the corresponding coefficient of *s* support vector $\alpha^* = (\alpha_1^*, \alpha_2^*, \ldots, \alpha_s^*)$ is not 0. We can also derive the weight coefficients of C_Z and C_E. The classification decision function of the testing image is expressed as follows:

$$f(B_{Zj}, B_{Ej}) = \text{sgn}\{\sum_{i=1}^{s} y_i \alpha_i [C_Z K(B_{Zi}, B_{Zj}) + C_E K(B_{Ei}, B_{Ej})] + b^*\} \quad (9)$$

In Formula (9), the classification threshold of b^* can be obtained by using any support vector.

4. Result Analysis And Discussion

Experimental simulations were conducted by using five public data sets, namely, MSRcv2[21], Caltech101[22], Scene15[10], Indoor67[23], and UIUC-Sport[24], to verify the validity of the method used in this study. From all of the image data sets, image blocks of different dimensions are extracted by crossing, with step length fixed at 8 pixels and dimensions being 16 × 16, 32 × 32, and 64 × 64, and are described by scale-invariant feature transform (SIFT) operators. At low-rank sparse decomposition, the parameter is $\Upsilon = \sqrt{MAX(T, D)}$, where *T* and *D* stand for the number and dimensions of the SIFT feature in the local feature matrix, respectively. *K*-means clustering is conducted on a randomly selected subset in the set of the low-rank and sparse features of all of the training images, with the number of *K* being set at 1,500 and the number of contiguous words in LLC set at 5. In the process of multiple kernel learning, the radical basis function is the linear kernel function, the penalty coefficient is 10, the "one versus one" decomposition policy is used in multi-class classification, and the maximum iterations of the outer and inner loops are set at 200.

4.1 Results of MSRcv2

The data set of MSRcv2 contains 15 kinds of objects, each having 30 images. Nine objects are selected for the experiment, namely, cars, planes, bicycles, human faces, books, cattle, and sheep, as shown in Figure.2. Using the Universal Testing Protocol[21], 15 images of each object are randomly selected as the training set; the remaining images are utilized as the testing set. The selections are repeated 10 times, and the average value of 10 times is selected as the experimental result.

The comparison of the results of the method proposed in this study and those of other new image classification methods in MSRcv2 are shown in Table.1. The proposed method outperforms all of the other contrast methods, with the highest average classification accuracy. The method used in Reference [26] has the best classification performance in MSRcv2. The results shown in Table.1

Figure 2. Sample images of the MSRCv2 data set

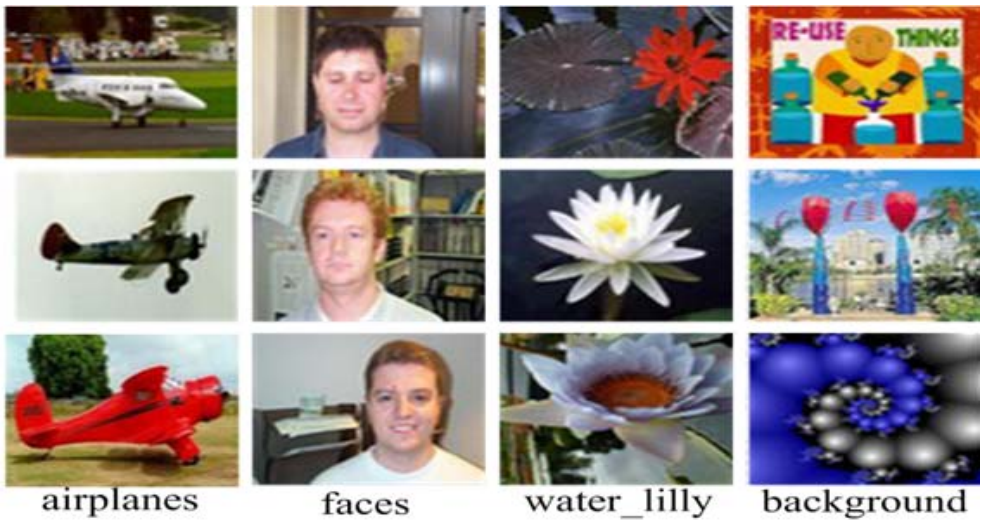

Figure 3. Sample images of the Caltech101 data set

Classification algorithm	Accuracy ± standard deviation (%)
Zhang and Chen [21]	80.4 ± 2.5
Sawaree [25]	80.00
Su [26]	90.7 ± 1.8
This paper	**90.71 ± 2.31**

Table 1. Performance comparison by classification in the MSRcv2 data set

The proposed method outperforms all of the other contrast methods, with the highest average classification accuracy. The method used in Reference [26] has the best classification performance in MSRcv2. The results shown in Table.1 reveal that the method proposed in this study achieves comparable classification performance, which shows the effectiveness of the method.

4.2 Results of Caltech101

The data set of Caltech101 contains 101 kinds of objects

and 1 background, with the number of images for each object ranging from 40 to 800, as shown in Figure.3. Using the Universal Testing Protocol[22], 30 images of each object are randomly selected as the training set; the remaining images are utilized as the testing set. The selections are repeated 10 times, and the average value of 10 times is selected as the experimental result.

The comparison of the results of the method proposed in this study and those of other new image classification methods in Caltech101 are shown in Table.2. The results shown in Table.2 reveal that the proposed method outperforms all of the other contrast methods, with the highest average classification accuracy of 8.31%, which is higher than the baseline algorithm used in Reference [22]. In References [1,2,4–9], classification performance is enhanced by improving the coding algorithms of the local features, among which References [6] and [8] have been published in top international academic journals in recent years. Table.2 shows that the method proposed in this study outperforms the methods presented in the eight reference papers by 11.77%, 2.71%, 2.47%, 1.70%, 1.44%, 2.68%, 0.89%, and 2.51%, respectively. This finding indicates the effectiveness of the proposed method in improving the BoF model at the feature extraction stage. If the method proposed in this study is combined with other methods, the classification performance will further improve. The methods presented in References [12] and [13] are the most similar to the method proposed in this study; however, the proposed method outperforms them by 0.23% and 0.08%, respectively, which further shows the effectiveness of the method.

Classification algorithm	Accuracy ± standard deviation (%)	Classification algorithm	Accuracy ± standard deviation (%)
Van Gemert [1]	64.14 ± 1.18	Yang and Yu [2]	73.20 ± 0.54
Wang and Yang [4]	73.44	Liu [5]	74.21 ± 0.81
Wang and Feng [6	74.47 ± 0.46	Shabou [7]	73.23 ± 0.81
Zhang and Ghanem [8]	75.02 ± 0.74	Wu [9]	73.4 ± 1.2
Zhang and Liu [12]	75.68 ± 0.89	Zhang and Ma [13]	75.83 ± 0.71
Griffin [22] 67.6	Jia [27]	75.3 ± 0.7	
This paper	**75.91 ± 1.22**		

Table 2. Performance comparison by classification in the Caltech101 data set

4.3 Results of Scene 15

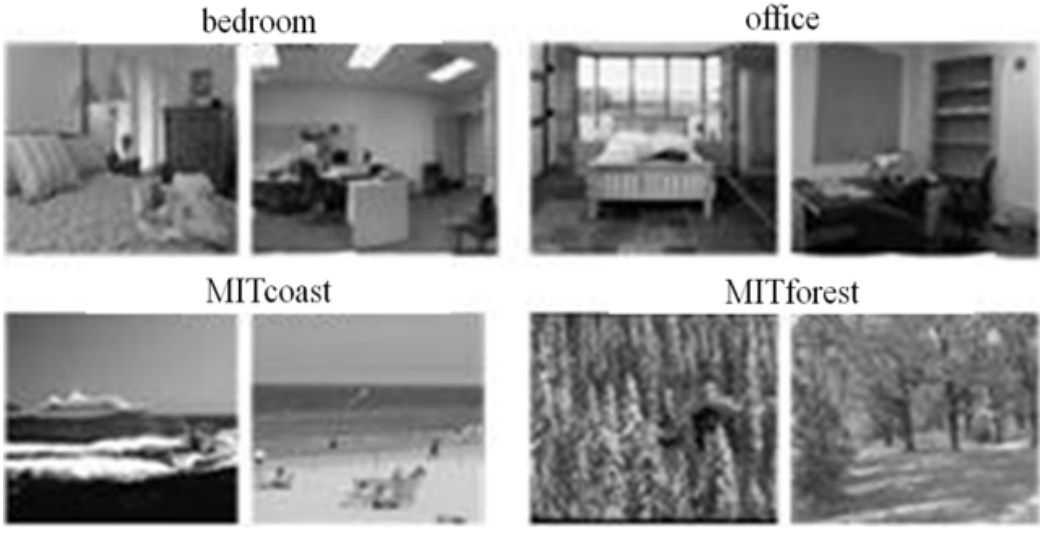

Figure 4. Sample images of the Scene15 data set

The data set of Scene15 contains 15 kinds of natural and man-made scenes, with the number of images for each scene ranging from 200 to 400, as shown in Figure.4. Using the Universal Testing Protocol[10], 100 images of each kind are randomly selected as the training set; the remaining images are utilized as the testing set. The selections are repeated 10 times, and the average value of 10 times is selected as the experimental result.

The comparison of the results of the method proposed in this study and those of other new image classification methods in Scene15 are shown in Table.3. Table.3 shows that the proposed method outperforms all of the other contrast methods, with the highest average classification accuracy of 1.96%, which is higher than the baseline algorithm used in Reference [10]. In References [1,4,5,7], classification performance is enhanced by improving the coding algorithms of the local features. Table.3 shows that the method proposed in this study outperforms the methods in the four reference papers by 6.69%, 1.86%, 0.66%, and 0.69%, respectively. This finding indicates the effectiveness of the method proposed in this study.

Classification algorithm	Accuracy ± standard deviation (%)	Classification algorithm	Accuracy ± standard deviation (%)
Van Gemert [1]	76.67 ± 0.39	Wang and Yang [4]	81.50 ± 0.87
Liu [5]	82.7 ± 0.39	Shabou [7]	82.67 ± 0.51
Lazebik [10]	81.40 ± 0.50	Yang and Newsam [28]	82.51 ± 0.43
This paper	**83.36 ± 0.49**		

Table 3. Performance comparison by classification in the Scene15 data set

4.4 Results of Indoo 67

Figure 5. Sample images of the Indoor67 data set

Classification algorithm	Accuracy ± standard deviation (%)	Classification algorithm	Accuracy ± standard deviation (%)
Quattoni [23]	26.5	Li and Su [29]	37.6
Bo [30]	41.8	This paper	**44.18**

Table 4. Performance comparison by classification in the Indoor67 data set

The data set of Indoor67 contains 67 kinds of indoor scenes, with the number of images for each scene

exceeding 100, as shown in Figure.5. Using the Universal Testing Protocol[23], 80 images of each kind are randomly selected as the training set; the remaining 20 images are utilized as the testing set.

The comparison of the results of the method proposed in this study and those of other new image classification methods in Indoor67 are shown in Table.4. Table.4 shows that the proposed method outperforms all of the other contrast methods, with the highest average classification accuracy of 17.68%, which is higher than the baseline algorithm used in Reference [23]. References [29] and [30] were published in top international academic journals in 2010 and 2011, respectively; the method proposed in this study outperforms them by 6.58% and 2.38%, respectively.

4.5 Results of UIUC-Sport

Figure 6. Sample images of the UIUC-Sport data set

Classification algorithm	Accuracy ± standard deviation (%)	Classification algorithm	Accuracy ± standard deviation (%)
Gao [3]	85.31 ± 0.51	Shabou [7]	87.23 ± 1.14
Bo [30]	85.7 ± 1.3	Li and Li [24]	73.4
Zhang and Liu [12]	86.69 ± 1.66	Zhang and Ma [13]	86.18 ± 1.21
This paper	**88.94 ± 0.96**		

Table 5. Performance comparison by classification in the UIUC-Sport data set

The data set of UIUC-Sport contains 8 kinds of sports, with the number of images for each kind of sport ranging from 137 to 2,500, as shown in Figure.6. Using the Universal Testing Protocol[24], 70 images of each kind are randomly selected as the training set and 60 images are utilized as the testing set. The selections are repeated 10 times, and the average value of 10 times is selected as the experimental result.

The comparison of the results of the method proposed in this study and those of other new image classification methods in UIUC-Sport are shown in Table.5. Table.5 shows that the proposed method outperforms all of the other contrast methods, with the highest average classification accuracy. In References [3,7], classification performance is enhanced by improving the coding algorithms of the local features, and Reference [3] has been recently published in a top international academic journal. Table.5 shows that the method proposed in this study outperforms the methods in the two reference papers by 3.63% and 1.71%, respectively. This finding indicates the effectiveness of the proposed method in improving the BoF model at the feature extraction stage. The methods in References [12,13] are most similar to that used in this study; however, the proposed method outperforms them by 2.25% and 2.76%, respectively, which further shows the effectiveness of the method.

5. Conclusion

BoF has been the most widely used method in image classification in recent years. In this study, a framework for image classification is established based on the BOF model. The framework clarifies the operation steps, operation object, and operation method in each step to achieve image classification. This study decomposes the local features of images to improve the representational features of the BoF model further. The following conclusions are obtained:

(1) In the proposed method, the local features of the images are decomposed as the low-rank and sparse parts, thus constructing and fusing the LR-BoF and S-BoF models through the SVM based on multiple kernel learning to achieve image classification.

(2) Parallel tests of five different public data sets, namely, MSRcv2, Caltech101, Scene15, Indoor67, and UIUC-Sport show that the proposed method achieves the highest average classification accuracy and is more effective for the improvement of the BoF model in local feature extraction, exhibiting a more robust and discriminative performance in image classification.

This study aims to improve image classification performance by constructing a new BoF model on the basis of feature decomposition. However, this method may lead to linear increase in feature dimensions, thus limiting its use in the classification of images of various types. Therefore, obtaining a more accurate and rapid image feature representation should be the direction of our future efforts.

Acknowledgements

The study was supported by the National Natural Science Foundation of China (61402192) and Major College Funding Project in Natural Science Foundation Projects of Jiangsu Province, China (15KJA460004)

References

[1] Van Gemert, J. C., Veenman, C. J., Smeulders, A. W., Geusebroek, J. M. (2010). Visual word ambiguity. *IEEE transactions on pattern analysis and machine intelligence*, 32(7) 1271-1283.

[2] Yang, J., Yu, K., Gong, Y., Huang, T. (2009). Linear spatial pyramid matching using sparse coding for image classification. *In: Computer Vision and Pattern Recognition(CVPR'2009),* p. 1794-1801. Miami, FL, USA: IEEE, June.2009.

[3] Gao, S., Tsang, W. H., Chia, L. T. (2013). Laplacian sparse coding, hypergraph laplacian sparse coding, and applications. *IEEE Transactions on Pattern Analysis & Machine Intelligence,* 35 (1) 92-104.

[4] Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., & Gong, Y. (2010). Locality-constrained linear coding for image classification. *In: Computer* Vision *and Pattern Recognition (CVPR 2010),* p. 3360-3367. San Francisco,USA: IEEE .November.2010.

[5] Liu, L., Wang, L., Liu, X. (2011). In defense of soft-assignment coding. *In: 2011 International Conference on Computer Vision (ICCV 2011)* , p. 2486-2493. Barcelona, Spain:IEEE. November.2011.

[6] Wang, Z., Feng, J., Yan, S., Xi, H. (2013). Linear distance coding for image classification. *IEEE Transactions on Image Processing*, *22* (2) 537-548.

[7] Shabou, A., LeBorgne, H. (2012). Locality-constrained and spatially regularized coding for scene categorization. *In: Computer Vision and Pattern Recognition (CVPR'2012),* p. 3618-3625. RI,USA: IEEE.June.2012.

[8] Zhang, T., Ghanem, B., Liu, S., Xu, C., Ahuja, N. (2013). Low-rank sparse coding for image classification. *In: I*EEE International Conference on Computer Vision(ICCV'2013), p. 281-288. Sydney, Australia:IEEE. December 2013.

[9] Wu, Z., Huang, Y., Wang, L., Tan, T. (2012). Group encoding of local features in image classification. *In: Pattern Recognition (ICPR'2012),* p. 1505-1508. Tsukuba Science City, Japan: Wikicfp.November.2012.

[10] Lazebnik, S., Schmid, C., Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'2006),* p.2169-2178. New York, USA:IEEE, June. 2006.

[11] Aybat, N. S., Goldfarb, D., Ma, S. (2014). Efficient algorithms for robust and stable principal component pursuit problems. *Computational Optimization and Applications*, 58.(1) 1-29.

[12] Zhang, C., Liu, J., Tian, Q., Xu, C., Lu, H., Ma, S. (2011). Image classification by non-negative sparse coding, low-rank and sparse decomposition. *In: Computer Vision and Pattern Recognition (CVPR'2011),* p. 1673-1680. Colorado, USA: IEEE, June.2011.

[13] Zhang, L., Ma, C. (2012). Low-rank, sparse matrix decomposition and group sparse coding for image classification. *In: 2012 19th IEEE International Conference on Image Processing* (ICIP'2012), pages 669-672.Orlando, Florida, USA: IEEE, Sep. 2012.

[14] Candès, E. J., Li, X., Ma, Y., Wright, J. (2011). Robust principal component analysis?. *Journal of the ACM*, *58*(3)11.

[15] Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., & Ma, Y. (2013). Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1)171-184.

[16] Boix, X., Roig, G., Leistner, C., Van Gool, L. (2012). Nested sparse quantization for efficient feature coding.

In: Computer Vision–ECCV. pages 744-758. Florence, Italy: Springer Berlin Heidelberg.October.2012.

[17] Shaban, A., Rabiee, H. R., Farajtabar, M., Ghazvininejad, M. (2013). From local similarity to global coding: An application to image classification. *In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR '2013),* pages 2794-2801.Portland,Oregon, USA:IEEE.June.2013.

[18] Najjar, A., Ogawa, T., Haseyama, M. (2015). Bregman pooling: feature-space local pooling for image classification. *International Journal of Multimedia Information Retrieval*, 4 (4) 247-259.

[19] Izquierdo-Verdiguier, E., Laparra, V., Gómez-Chova, L., Camps-Valls, G. (2013). Encoding invariances in remote sensing image classification with SVM. *IEEE Geoscience and Remote Sensing Letters*, *1*0 (5) 981-985.

[20] Ahmed, F., Al-Mamun, H. A., Bari, A. H., Hossain, E., Kwan, P. (2012). Classification of crops and weeds from digital images: A support vector machine approach. *Crop Protection*, 40,98-104.

[21] Zhang, Y., Chen, T. (2009). Efficient kernels for identifying unbounded-order spatial features. *In: Computer Vision and Pattern Recognition(CVPR'2009).* p. 1762-1769.Miami, FL, USA: IEEE, June.2009.

[22] Griffin, G., Holub, A., Perona, P. (2007). Caltech-256 object category dataset. *California Institute of Technology*.1-20.

[23] Quattoni,A.,Torralba,A.(2009). Recognizing indoor scenes. *In: Computer Vision and Pattern Recognition(CVPR'2009)*. pages 413-420.Miami, FL, USA: IEEE, June.2009.

[24] Li, L. J., Fei-Fei, L. (2007). What, where and who? classifying events by scene and object recognition. *In:* 2007 IEEE 11th International Conference on Computer Vision, p. 1-8. Rio de Janeiro. Brazil: IEEE. October. 2007.

[25] Savarese, S., Winn, J., Criminisi, A. (2006). Discriminative object class models of appearance and shape by correlatons. In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'2006)*,pages 2033-2040. New York, USA:IEEE, June. 2006.

[26] Su, Y., Jurie, F. (2011). Visual word disambiguation by semantic contexts. *In: 2011 International Conference on Computer Vision,*pages 311-318. Barcelona, Spain: IEEE. November.2011.

[27] Jia, Y., Huang, C., Darrell, T. (2012). Beyond spatial pyramids: Receptive field learning for pooled image features. *In: Computer Vision and Pattern Recognition (CVPR'2012),* p. 3370-3377.RI,USA: IEEE.June.2012.

[28] Yang, Y., Newsam, S. (2011). Spatial pyramid co-occurrence for image classification. *In: 2011 International Conference on Computer Vision* ,pages 1465-1472. Barcelona, Spain: IEEE. November.2011

[29] Li, L. J., Su, H., Fei-Fei, L., Xing, E. P. (2010). Object bank: A high-level image representation for scene classification & semantic feature sparsification. *In: Advances in neural information processing systems (NIPS'2010)*, pages 1378-1386. Vancouver, B.C., Canada: Neural Information Processing Systems. December .2010.

[30] Bo, L., Ren, X., Fox, D. (2011). Hierarchical matching pursuit for image classification: Architecture and fast algorithms. *In: Advances in neural information processing systems* (NIPS'2011),p. 2115-2123. Granada, Spain : Neural Information Processing Systems. December. 2011.