ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 134

2nd International Conference on Control, Automation, and Artificial Intelligence (CAAI 2017)

A New VLAD Method with Dense SIFT Selection Application in Image Classification

Zhi Qian, Qijun Hong, Gang Huang, Pingping Liu*, Yuanjie Yan and Min Xie

College of Computer Science and Technology, Jilin University, China

*Corresponding author

***Abstract*—Since Dense SIFT causes a long time spending during clustering due to an excessive order of magnitude, and its feature descriptors reserve excessive insignificant features, we present a new method that using SLIC to select descriptors to address this problem. Furthermore, when VLAD aggregates, the partial directions of feature vectors have the excessive data offset and still distorts after the dimension deduction treatment. Regarding such issue, the algorithm that possesses the optimized clustering descriptor with feature membership information called FS-VLAD is proposed. The algorithm adopts the principle of the fuzzy cost function with the smallest deviation regarding the quadratic sum of the neighbor clustering center to calculate the feature membership degree. After conducting classification test, the result demonstrates that in comparison with the mainstream Dense SIFT + VLAD classification model, the new methods could improve by around 15%, and possesses better generality.**

Keywords-image classification; features selection; SLIC; VLAD(key words)

I. INTRODUCTION

Recently, as the improvement of the Bag of Words (BOW) [12] method, Vectors of Locally Aggregated Descriptors (VLAD) [10, 13] has been very prevalent within the image classification domain. General classification method is to utilize Dense Scale-invariant feature transform (Dense SIFT) to extract the descriptors of the training set images, then apply the k-means method to divide the descriptors of each image into N clustering, and the clustering center is the word. The N clustering can be used to cluster VLAD feature vectors, and these feature vectors can be transferred to the corresponding classifier (SVM [19] in general) to train classifiers. [1] This method possesses the characteristics of simple generation procedures, and superior query performance, through which a relatively superior result can be achieved.

In this paper, two approaches are presented for these issues: 1). Using SLIC superpixel segmentation [2, 7] for the Dense SIFT descriptor set extracted, and lower the amount of the descriptors with insignificant features as well as the total sum of the descriptor set; 2). Improving the calculation procedures of VLAD, and eliminate the cumulative residuals produced from the VLAD feature generation process. In our opinion, each small area generated through SLIC superpixel segmentation basically only incorporates some similar features. However, multiple circumstances might occur for the features among different areas. Therefore, the descriptors between the two segmentation areas which incorporate features from both of the areas, are extremely significant that they can effectively represent the picture. As for VLAD, we have utilized the comprehensive evaluation method in fuzzy numbers, and calculated the feature membership degree as the weight coefficient that has demonstrated more precise distribution circumstances of feature vectors. This has solved the problem that VLAD directly calculates the offset of cumulative residuals of the closest clustering center.

The paper is organized as follows: In Section 2, some related works of image classification based on Dense SIFT (or other SIFT descriptor) and VLAD are briefly described. In Section 3, the detailed principles and methods regarding the descriptor selection through SLIC boundary segmentation are introduced. In Section 4, the feature membership degree calculation of FS-VLAD and generation process of clustering descriptor are introduced. The experimental results are presented in Section 5. Some conclusions on the results are discussed in section 6.

II. RELATED WORKS

This section has reviewed some methods in solving the problem of massive local feature descriptors by other scholars, as well as the optimized methods regarding VLAD.

Regarding the issue of the excessive amounts of Dense SIFT descriptors, Vedaldi and Fulkerson[1] has utilized the randperm function of Matlab to disorganize the descriptor matrix, and extracted the first n descriptions after disorganization and conducted resequencing (n is the amount of descriptors in each image that has been preset) in the VLFEAT computer visual library they established, in order to guarantee the relative positions of these n descriptors are not changed.[8] Such method can reduce the amounts of descriptors, meanwhile reserve the descriptors in each area of the image in relatively even way, however it can still cause information loss. Moreover, regarding the issue that massive insignificant Dense SIFT descriptors occupy a large proportion in the keywords with low differentiation degree in feature vectors, Vedaldi and Fulkerson [1] deem that SVM will reduce the weights for these keywords during the learning process. However, considering the differences among a variety of training sets, this method is not highly reliable.

In terms of the cumulative residuals of VLAD, Delhumeau and Gosselin [14] proposed two combination approaches which are Residual normalization (RN) and Local coordinate system (LCS). First of all, they conducted the direct normalization for all of the descriptor cumulative residuals, see in (1).

$$v_i = \sum_{x:NN(x)=\mu_i} \frac{x-\mu_i}{\|x-\mu_i\|} \quad (1)$$

In their view, a better suitability could be achieved through independently adjusting the coordinate system of each visual word. For the word that sequenced as number i (i=1…k), the spin matrix Qi can be achieved through the training descriptive character that has mapped into that word (for specific data set). Thereafter, they cluster the k-dimensional spin matrix into VLAD, and apply the k-dimensional spin matrix into the normalization cumulative residual vector, see in (2).

$$v_i = \sum_{x:NN(x)=\mu_i} Q_i \frac{x-\mu_i}{\|x-\mu_i\|} \quad (2)$$

Their measures do optimize the precision in some specific data set, however, in order to obtain the spin matrix, PCA for each Voronoi unit of the feature space is required. Such excessively early normalization and locally conducting PCA can cause the overall information loss and the deviation can be further magnified after the final dimension deduction.

S Husain and M Bober [15] have proposed a kind of vision descriptive character based on permutation, the Robust Visual Descriptor (RVD). In the method of RVD, each feature is quantified into k-neighboring visual words, these k-neighboring visual words are sequenced in a numerical order and the cumulative residuals are calculated. Different cumulative residuals are permuted into each visual word, and $\{rv_1, rv_2, rv_i, rv_k\}$ can be achieved through clustering. Finally, the weight $\{\omega_1, \ldots, \omega_i, \omega_k\}$ can be determined according to the permutation information $\{k, k-1, \ldots, 1\}$. Thereafter, the cumulative residuals of all the features in words are clustered, among which the cumulative residual for the word can be expressed as:

$$v_i = \omega_1 * rv_1 + \omega_2 * rv_2 + \cdots + \omega_k * rv_k \quad (3)$$

III. DESCRIPTOR SELECTION

A. *The Characteristics of Dense SIFT*

Dense SIFT is derived from the Scale-Invariant Feature Transform (SIFT) [9], and the most significant difference between them is that Dense SIFT assumes all the significant points are evenly distributed. Therefore, the selection of keywords in all areas of the image is dense and standardized. The significance of SIFT-like descriptor can be demonstrated by the difference of each direction that passes through its Histogram of Oriented Gradients. If the difference is obvious between the feature's main directions and its supporting directions, the descriptor is significant. Otherwise, the descriptor is insignificant. It can be found that most of the descriptors generated by Dense SIFT are insignificant. The differentiation degree is not high for the words generated by massive insignificant descriptor clustering. As shown in Figure I.

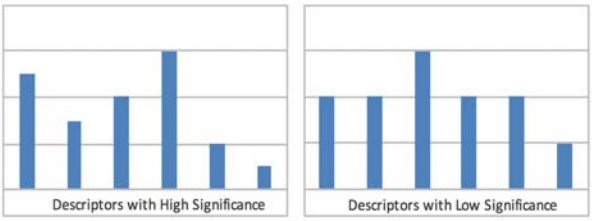

FIGURE I. HISTOGRAM OF ORIENTED GRADIENTS FOR DESCRIPTORS WITH DIFFERENT SIGNIFICANCE. X AXIS SHOWS TOP FIVE MAIN DIMENSIONS OF DESCRIPTORS AND Y AXIS SHOWS THE VALUE OF THEM.

B. *The Principle for Descriptor Selection through Dividing SLIC into Areas*

The process of Simple Linear Iterative Clustering (SLIC) is as follows: First, converting the color images into CIELAB color space and the 5-dimensional feature vector $C_i = [l_i a_i b_i x_i y_i]$ under the XY coordinates; then constructing the standards of distance measurement in terms of the 5-dimensional feature vector C_i, and conducting local clustering of image pixels. [2, 7] This method can generate compact and almost evenly distributed superpixels, which is highly recognized in aspects of operation speed, object contour preservation, and the superpixel shape generated, etc., and relatively meets the expectation in terms of the segmentation effects.

Each pixel within every single superpixel area after the SLIC segmentation usually possesses similar feature of color. Moreover, between the nearby areas, there are two circumstances depending on the difference of reasons for the area segmentation [11]:

- The area segmented due to the excessively large color distances. In such circumstance, the gradient of pixels within the two areas is usually large, and in general, the area boundary is also the actual boundary of objects in the image.
- The area segmented due to the excessively large space distances. In such circumstance, the gradient of pixels within the two areas is usually small, and the two areas belong to the same entity in the picture.

Due to such characteristics of the SLIC area boundary, the Dense SIFT descriptors can be sorted into two categories:

- For the descriptors on the area boundary caused by circumstance a, since the information from the two edges of the boundary is different, the descriptors on the boundary possess highly significant main directions with high differentiation degrees.
- For the descriptors massively existing within the area and the descriptors on the boundary caused by circumstance b, their main directions are not significant, and the differentiation degree is low.

Therefore, if only the descriptors on SLIC segmentation area are preserved and those within massive area are abandoned, it can be achieved simultaneously that:

- To reduce the proportion of descriptors with low differentiation degree, meanwhile to preserve the image information they convey, which can enable them to play a proper role in the feature vectors.

- To preserve the descriptors with high differentiation degree from being abandoned. Through this method, both of the precision and disposal speed in terms of the image segmentation can be improved.

As shown in Figure II, under the same circumstances that the number of descriptors is 256, it can be observed that the descriptors after the SLIC selection are more evenly distributed than those having not went through the SLIC selection process, and the amount of descriptor is less on the background, and the number of descriptors in areas with obvious features (e.g. the contour and eyes of the cat) is more. Figure III has briefly demonstrated the proportion of the keywords generated by those two categories of descriptors.

FIGURE II. NORMAL DENSE SIFT DESCRIPTOR DISTRIBUTION (LEFT) AND DESCRIPTOR DISTRIBUTION AFTER SLIC SELECTION (RIGHT).

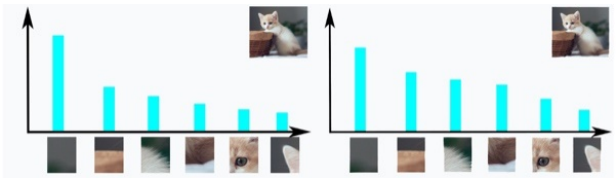

FIGURE III. THE TOTAL PROPORTION OF PARTIAL KEYWORD GENERATED BY NORMAL DENSE SIFT (LEFT) AND THE TOTAL PROPORTION OF PARTIAL KEYWORDS AFTER SLIC SELECTION(RIGHT). X AXIS SHOWS TOP SIX MAIN DIMENSIONS OF FEATURE AND Y AXIS SHOWS THE VALUE OF THEM.

C. *Detailed Procedures for Feature Selection*

Thus, we proposed the detailed selection procedures based on VLFEAT computer visual library.

- Step1: Retrieve the coordinates of the pixels from SLIC index value S.

- Step2: Construct KD-tree [18] for all the coordinates.

- Step3: Introduce Dense SIFT feature for queries, and acquires the boundary pixel distance t.

- Step4: Compare the distance t with the radius of the feature area.

- Step5: If the distance is larger than the radius, it must be deleted, otherwise repeat step 3 query Dense SIFT feature points selection.

The procedure is demonstrated as in Algorithm 1.

Algorithm 1. Features filtrate by SLIC

INPUT: SLIC rand index s and Dense SIFT features
OUPUT: filtrate feature

```
feature = select_sift_by_slic(height, s, features)
// Retrieve the coordinates
for i ← 1 to size(s,1) do
    coordinates(1,i) ← mod(s(i,1) , height)
    coordinates(2,i) ← fix(s(i,1) / height) + 1
end for
//Build kd-tree
kdtree ← kdtreebuild(coordinates)
//del descriptors
for i ← 1 to size(s,1) do
    [index,t] ← kdtreequery(coordinates)
    if t > features.frame(3,i)
        Then features ← 0
end for
Del features which are equal 0
return features
```

IV. OPTIMIZATION OF THE CLUSTERING DESCRIPTOR

A. *Feature Membership Degree*

In the original VLAD process, it is similar with the BOW process that the codebook $C = \{\mu_1, \ldots, \mu_k\}$ is achieved first through clustering algorithms (e.g. K-means), then the descriptor in Dense SIFT local descriptor set $X = \{x_1, \ldots, x_t\}$ can be distributed into the closest clustering center through (4). [10].

$$\mathrm{NN}(x_t) = \arg \min_{u \in C} \|x\text{-}\mu_i\| \tag{4}$$

Thereafter, each feature vector needs to be clustered, for each quantified index $i \in [1, \ldots, k]$ through cumulative residual vectors (i.e. the differences between the descriptive character and its distributed clustering center), the d-dimensional sub vector v_i can be achieved according to (5).

$$v_i = \sum_{x:NN(x_t)=\mu_i} (x\text{-}\mu_i) \tag{5}$$

For the D-dimensional vector of VLAD, $D = K * d$ where d is the dimension amount of d-dimensional local descriptors, and PCA dimension deduction is usually needed for the d-dimensional Dense SIFT descriptors. At last, after the combination of the v_i sub vector, the L2 normalization is conducted, and $V := \frac{v}{\|v\|}$.

VLAD is a compact description generated after the clustering of local features such as SIFT, etc. Although it is the aggregated sum of the cumulative residuals of K clustering centers and merged with location information, it still cannot manifest the attribution weight between the feature and clustering center. The SIFT local descriptors of an image directly accumulate its all the cumulative residuals of the

closest clustering center, which might cause a shift of the gravity center, and might still cause partial distortions after dimension deduction treatment.

Membership degree is a concept in the evaluation functions of the fuzzy mathematics. [16] The fuzzy comprehensive evaluation is an effective multifactor decision-making method that can provide comprehensive evaluation towards the objects influenced by multiple factors, and its characteristics are that its evaluation result is not absolutely positive or negative, but shown through a fuzzy set. Therefore, different weight values can be divided according to the distances from the local descriptor to the clustering center, which is called feature membership weight value. This has offered VLAD a method of reformulation, making VLAD possess understandings and conceptions regarding some of the fuzzy words.

The feature membership weight calculates based on the cost function of the fuzzy clustering [17], in which the deviation is required as the minimum regarding the quadratic sum of the vector and clustering center during the clustering process, the cost function is shown as (6).

$$J_b = \sum_{i=1}^{k}\sum_{j=1}^{n}\left[f_i(S_j)\right]^r \left\|S_j\text{-}\mu_i\right\|^2 \tag{6}$$

The smaller the cost function value is, the smaller the deviation is. In the formula, k is the clustering amount, S_j represents the jth local descriptive vector, $f_i(S_j)$ is the membership degree for the feature vector S_j to the ith clustering, r is constant that is larger than 1, which can control the extent of the fuzzy clustering. In the experiment, the value of r is 2, and the aggregate sum of the membership degree for each SIFT feature vector to every clustering is determined as 1, such as shown in (7).

$$\sum_{i=1}^{k} f_i(S_i) = 1, j = 1,2,\cdots,n \tag{7}$$

Since the cost function J_b has its minimum value, under the membership constraint of Ep.6, by determining the value of partial derivative for J_b to $f_i(S_i)$ and S_j is 0, a prerequisite can be achieved as in (8), and the cluster center can be acquired.

$$\mu_i = \frac{\sum_{j=1}^{n}[f_i(s_j)]^r s_j}{\sum_{j=1}^{n}[f_i(s_j)]^r}, \mathrm{i} = 1,2,\cdots,k \tag{8}$$

Therefore, the membership degree $f_i(S_j)$ for S_j in the ith clustering can be achieved through calculations in (9).

$$f_i(S_j) = \frac{\left(1/\|s_k-\mu_i\|^2\right)^{\frac{1}{r-1}}}{\sum_{h=1}^{k}\left(1/\|s_j-\mu_h\|^2\right)^{\frac{1}{r-1}}}, j = 1,2,\cdots,n,\ i = 1,2,\cdots,k \tag{9}$$

B. *The Generation of FS-VLAD*

The procedures for generating FS-VLAD features are shown as the following:

- Step1: Initialize the FS-VLAD vector.
- Step2: Conduct k-neighbor queries regarding the Dense SIFT descriptors, and find the kth clustering center.
- Step3: Calculate the distance with the kth clustering center for descriptors, and further calculate the feature membership weight value.
- Step4: Calculate the difference between the descriptor S and the kth clustering center through the feature membership weight value.

The FS-VLAD generated also possesses the dimension of D*K, where D is the dimension of Dense SIFT.As the same as VLAD, the dimension of the initial FS-VLAD is relatively high, and it needs dimension deduction treatment prior to the index. After applying the PCA linear dimension deduction, through the comparison based on the experiment, the precision and average precision ration is higher for FS-VLAD comparing to VLAD with the same setting.

The detailed procedures of the optimized clustering descriptor algorithm are demonstrated as the Algorithm 2:

Algorithm 2 Generate the FS-VLAD

INPUT: The codebook set generated by k-means C ={$\mu_1,\ldots,\mu_k$}, and the local descriptor set of Dense SIFT S=$s_1,\ldots, s_1$

OUPUT: FS-VLAD features

1 feature=FS-VLAD (clustering center k, features)
2 for i ←1 to k do
3 $v_i = 0_d$ // init v_i
4 end for
5 for i ← 1 to t do
6 $\mathrm{index}_{\mathrm{word}[1,\ldots,\mathrm{m}]} = \arg\min_{u \in C} \|\mathrm{S}_i\text{-}\mu_j\|$ // get the m words
7 Obtain feature membership weight $f_i(S_j)$ by (9)
8 $v_i = \sum \frac{1}{f_i(S)}(S\text{-}\mu_i)$
9 $V := (v_1,\cdots,v_i,\cdots,v_k)$
10 Apply L2 Normalization for V
11 return FS-VLAD feature

V. THE EXPERIMENT RESULTS AND ANALYSIS

A. *The Introduction of the Dataset and Evaluation Method*

We utilized the Pascal Visual Object Classes 2007 (VOC07) image set [3], Caltech 101 image set [4], Indoor Scene Recognition (Scene67) image set[5], and the Flickr Material Dataset (FMD) image set[6] in order to conduct experiments, and some comparisons are also made in terms of the improvement method based on Dense SIFT and VLAD. VOC07 image set incorporates four main categories and 20 small species with 9,963 images in total, and the purpose of this image set is to challenge the target recognition capabilities of the classification model under real scenarios. The Caltech 101 image set includes 102 categories with 9,144 images in total, and it's used for testing the recognition capabilities of the classification model towards different targets. Scene67 focuses

on the recognition under indoor scenarios with 67 categories and 15,620 images. FMD addresses the recognition towards 10 kinds of natural and artificial materials with 1,000 images. Though the verification of these four kinds of database, the high efficiency and wide applicability of our method can be proved.

The evaluation index for the target classification performance is Acc (accuracy) and average precision rate mAP (Mean Average Precision). According to different dataset characteristics, we utilized mAP as the evaluation standard of VOC07 and Acc as the evaluation standard of the other three datasets.

$$acc = \frac{\text{The sum of classification precision rates}}{Total\ number\ of\ the\ image\ categories} * 100\% \quad (10)$$

$mAP = \int_0^1 p(R)dR$, $p(R)$ is the precision rate for the system when the recalling rate is R.

$$mAP = \int_0^1 p(R)dR \quad (11)$$

B. The Selection of Dense SIFT through SLIC

As it has been proposed in the paper that the way to preserve only the descriptors at the SLIC segmentation areas and abandon the descriptors within the massive areas can simultaneously improve the precision rate and disposal speed of the image classification, we will conduct comparisons of classification precision rates between the VLAD feature of descriptor clustering after the SLIC selection and VLAD feature of descriptor clustering without the SLIC selection, and the experiment has been carried out with 4 datasets and achieved ideal results. The experiment results are shown in the table I, among which, REGIONSIZE is the starting size of the superpixels and REGULARIZER is the trades-off appearance for spatial regularity when clustering a larger value results in more spatial regularization.

Through Table I, it can be observed that when RegionSize = 30 and Regularizer =1, the precision rate is the highest, an act to decrease or increase the area of a single segmentation will decrease the classification precision. This is due to the reason that expanding the area might cause essential information loss, and an excessively small area might cause an excessive amount of the low-quality descriptors due to an excessive number of segmentation areas, which cannot meet the expected purpose. It should be noticed that such parameter is based on the image pixels of the training sets. The pixels from the four image sets used in this experiment are all within the range of 200*300 and 300*500.

TABLE I. THE CLASSFICATION PRECISION RATES FOR FEATURE POINT SELECTION UNDER DIFFEWRENT SLIC PARAMENTERS

Dataset		**VOC07(mAP)**	**Caltech 101(Acc)**	**Scene 67(Acc)**	**FMD(Acc)**
			VLAD + aug.		
		54.66%	**78.68%**	**53.29%**	**49.40%**
RegionSize	**Regularizer**		**VLAD+SLIC+ aug**		
100	10	56.26%	80.23%	56.11%	52.23%
100	1	56.63%	80.54%	56.42%	52.55%
100	0.1	56.65%	80.56%	56.44%	52.56%
50	10	56.82%	81.34%	57.01%	52.94%
50	1	57.02%	81.47%	57.22%	53.07%
50	0.1	57.01%	81.45%	57.22%	53.09%
30	10	57.56%	82.24%	57.62%	53.21%
30	1	57.72%	82.36%	57.73%	53.57%
30	0.1	57.77%	82.39%	57.75%	53.58%
10	10	57.27%	82.11%	57.54%	53.15%
10	1	57.45%	82.26%	57.68%	53.51%
10	0.1	57.48%	82.22%	57.70%	53.52%

C. The Comparison Experiment of VLAD

We carried out comparisons between the clustering descriptor FS-VLAD features after optimization and the VLAD features without optimization, and the experiments carried out on four datasets have all achieved ideal results.

Constant m represents the amount of a single Dense SIFT vector that belongs to the clustering. The experiment has found that when m is excessively large, the time spending for obtaining clustering is long, and the clustering excessively far with the feature vectors makes no sense since the membership degree weight value is too small, and an excessively small value of m cannot demonstrate the actual membership relations between the feature and clustering. As it can be shown in the data of Table II, the optimal performance is achieved when $m = 3$.

TABLE II. THE EFFECT OF DIFFERENT FUZZY CLUSTERING NUMBERS M ON DIFFERENT NUMBERS OF WORDS

K	VOC07(mAP) m = 3 m = 4	Caltech 101 m = 3 m = 4	Scene 67 m = 3 m = 4	FMD m = 3 m = 4
32	62.11 62.08	81.11 81.13	58.06 58.06	55.25 55.22
128	64.33 64.33	84.11 84.12	59.88 59.83	57.23 57.22
256	65.97 65.89	84.89 84.88	61.45 61.48	59.86 59.87

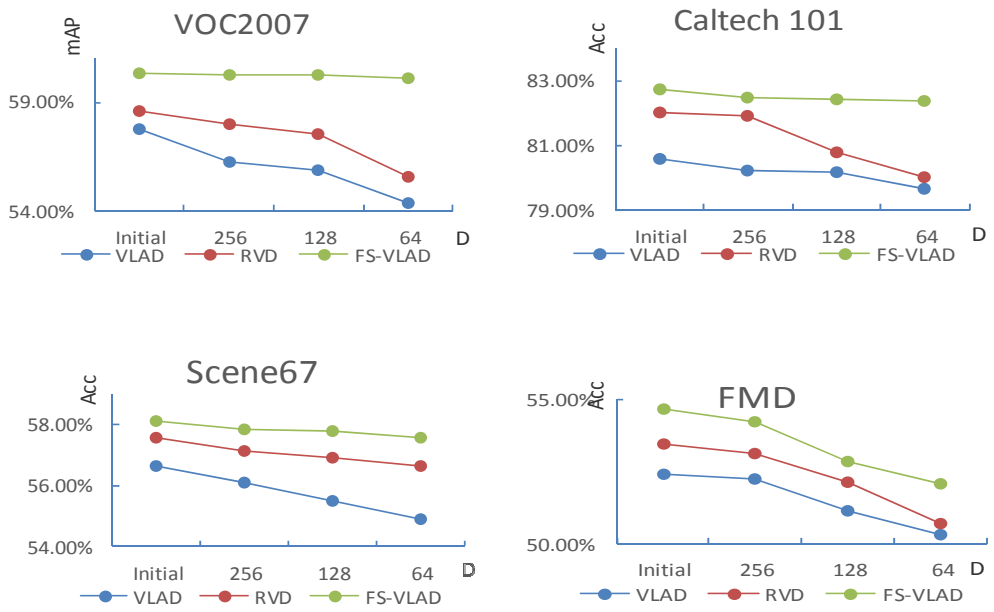

FIGURE IV. THE COMPARISONS OF DIFFERENT DEGREE OF DIMENSION DEDUCTION FROM THE UPPER LEFT TO LOWER RIGHT DIRECTION.

D. The Comparison Experiment after Integration

The four methods with the same parameter setting as much as possible have been utilized in the comprehensive comparison experiment, and the dimension of 256 has been attributed for each clustering descriptor after the dimension deduction. It can be observed from the comparison of the four image sets from Table III that, the FS-VLAD+SLIC selection method has improved by 15% in average based on the original VLAD method. Therefore, it can be proved that FS-VLAD+SLIC selection method possesses the significant advantages of high efficiency and strong applicability in the domain of image classification.

TABLE III. THE COMPARISON OF THE COMPREHENSIVE ALGORITHM AFTER COMBINING SLIC AREA SEGMENTATION SELECTION

	VOC07 (mAP)	Caltech 101	Scene 67	FMD
BOW	47.48%	70.11%	50.81%	45.77%
VLAD	54.66%	78.68%	53.29%	49.40%
RVD+SLIC	58.73%	80.29%	56.03%	51.11%
FS-VLAD +SLIC	**62.36%**	**83.10%**	**58.87%**	**54.22%**

VI. CONCLUSIONS

In the paper, we have proposed the method of using the superpixel SLIC algorithm to select Dense SIFT feature points, as well as the method of further optimizations for clustering descriptor VLAD to generate FS-VLAD feature vectors with feature membership degree. We have conducted experiments based on the datasets, such as VOC07, Caltech101, Scene 67, FMD, etc. by applying our method, and has proved that among the image classification models, there has occurred a significant improvement by applying our model comparing to the original model.

Future Works:

- Conduct rankings in terms of precision rates for the image segmentation area, and make the selection more effective.
- Consider the stratification selection for the feature points which can enhance the efficiency for the clustering conducted later.
- When generating FS-VLAD clustering vectors, design and distribute more suitable calculation methods in practice regarding different data.

REFERENCES

[1] Vedaldi, Andrea, and Brian Fulkerson. "VLFeat: An open and portable library of computer vision algorithms." Proceedings of the 18th ACM international conference on Multimedia. ACM, 2010.

[2] Achanta, Radhakrishna, et al. "SLIC superpixels compared to state-of-the-art superpixel methods." IEEE transactions on pattern analysis and machine intelligence 34.11 (2012): 2274-2282.

[3] Everingham, Mark, et al. "The PASCAL visual object classes challenge 2007 (VOC2007) results."

[4] Fe-Fei, Li. "A Bayesian approach to unsupervised one-shot learning of object categories." Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on. IEEE, 2003.

[5] Quattoni A, Torralba A. Recognizing indoor scenes[C]//Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009: 413-420.

[6] Sharan, Lavanya, Ruth Rosenholtz, and Edward Adelson. "Material perception: What can you see in a brief glance?." Journal of Vision 9.8 (2009): 784-784.

[7] Achanta, Radhakrishna, et al. Slic superpixels. No. EPFL-REPORT-149300. 2010.

[8] Bosch, Anna, Andrew Zisserman, and Xavier Munoz. "Image classification using random forests and ferns." Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on. IEEE, 2007.

[9] Lowe, David G. "Distinctive image features from scale-invariant keypoints." International journal of computer vision 60.2 (2004): 91-110.

[10] Arandjelovic, Relja, and Andrew Zisserman. "All about VLAD." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013.

[11] B. Fulkerson, A. Vedaldi, and S. Soatto, “Class Segmentation and Object Localization with Superpixel Neighborhoods,” Proc. IEEE Int’l Conf. Computer Vision, 2009.

[12] Sivic, Josef, and Andrew Zisserman. "Video google: A text retrieval approach to object matching in videos." iccv. Vol. 2. No. 1470. 2003.

[13] Jégou, Hervé, et al. "Aggregating local descriptors into a compact image representation." Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010.

[14] Delhumeau, Jonathan, et al. "Revisiting the VLAD image representation." Proceedings of the 21st ACM international conference on Multimedia. ACM, 2013.

[15] Husain, Syed, and Miroslaw Bober. "Robust and scalable aggregation of local features for ultra large-scale retrieval." Image Processing (ICIP), 2014 IEEE International Conference on. IEEE, 2014.

[16] Zadeh, Lotfi A. "Fuzzy sets." Information and control 8.3 (1965): 338-353.

[17] Xie, Xuanli Lisa, and Gerardo Beni. "A validity measure for fuzzy clustering." IEEE Transactions on pattern analysis and machine intelligence 13.8 (1991): 841-847.

[18] Muja, Marius, and David G. Lowe. "Fast approximate nearest neighbors with automatic algorithm configuration." VISAPP (1) 2.331-340 (2009): 2.

[19] Chapelle, Olivier, Patrick Haffner, and Vladimir N. Vapnik. "Support vector machines for histogram-based image classification." IEEE transactions on Neural Networks 10.5 (1999): 1055.