EDGE-BASED LOCALLY AGGREGATED DESCRIPTORS FOR IMAGE CLUSTERING

Yang Dong, Dazhao Fan*, Qiuhe Ma, Song Ji, Rong Lei

Zhengzhou Institute of Surveying and Mapping, Zhengzhou, China, wenku34@163.com

Commission III, ICWG II/III

KEY WORDS: Edge feature points, Local feature aggregation, Global image descriptors, Image clustering

ABSTRACT:

The current global image descriptors are mostly obtained by using the local image features aggregation, which fail to take full account of the details of the image, resulting in the loss of the semantic content information. It cannot be well used to make a good distinction between the high similarity images. In this paper, a new method of image representation, which can express the whole semantics and detail features of the image, is proposed by combining the edge features of the image. It is used to make a global description of the images and then clustering. The experimental results show that the proposed method is capable of clustering of the similarity images with high accuracy and low error rate.

1. INTRODUCTION

Clustering between images is an important step in the process of remote sensing data processing and analysis. It has been widely used in the fields of three-dimensional reconstruction, pattern recognition, object analysis and so on. The traditional clustering method mainly consists of two steps: one is generating the global descriptor of the image; another is clustering the image descriptors. It can be seen from that the global descriptor of the image directly affects the accuracy of subsequent processing. It is worth exploring that how to generate a good global image descriptor.

The global descriptor is the overall description of the image, which is usually obtained by aggregating the local image descriptors. For example, using the bag-of-words (BoW) model to aggregating the local image descriptors, which is simple and effective, is widely used in research and analysis (Csurka et al., 2004). Similarly, serval feasible and effective global description approaches for the image were developed, as referred in (Krizhevsky et al., 2012; Lazebnik et al., 2006; Perronnin et al., 2010; Russakovsky et al., 2012; Wang et al., 2010; Yang et al., 2009). With the help of the above global description approaches, the clustering accuracy of the images was significantly improved. However, the above approaches took only characterize the original image to some extent, lacks the semantic information of the original image. Therefore, it is still a hot research field that how to generate the global descriptor with good properties, make the image coding contain more semantic information, and have more obviously inter-class features and better in-class descriptions.

In this paper, aiming at the semantic information of the image, an edge-based local feature aggregation algorithm is proposed. Firstly, the edge of the image is extracted, and the global descriptor u_1 of the image is generated by using the edge point descriptors, which makes the descriptor contain strong semantic and detail description information. At the same time, the global image description u_2 is generated by the low-resolution image of the image pyramid. Finally, two descriptors are combined to obtain the integrated descriptor with both the image scene information and the strong detail information. Considering the semantic information contained in the image during the generation of the image synthesis descriptor, the image has good in-class description ability and inter-class discrimination ability, which improves the accuracy of the subsequent image clustering results.

2. GLOBAL IMAGE DESCRIPTORS

2.1 Global image descriptors based on BoW model

BoW model comes from text classification technology. The core idea is to use the image as the document object and use the local feature contained in the image as the word, so as to use a set of feature sets to describe the whole image. The general procedure for generating global image descriptor using BoW can be described as follows:

1) Performing image local feature extraction to get local feature set M;
2) Clustering the feature set $\Phi = \{M_1, M_2 \cdots M_i \cdots M_n\}$ of the image n to obtain m cluster centers, and taking each cluster center as a visual vocabulary φ to form a visual dictionary $\Psi = \{\varphi_1, \varphi_2 \cdots \varphi_i \cdots \varphi_m\}$;
3) Mapping each feature of each image feature set M to a word in the visual dictionary Ψ, and accumulating the number of times so that a word vector u of m dimension can be generated from the sum, that is, the global descriptor of the image.

It can be seen from the above process that the generation of the visual dictionary Ψ is an important step in BoW model. The generation of the visual dictionary directly affects the quality of the subsequent image clustering. In general, k-means clustering algorithm can be used to cluster the image features to generate visual dictionaries.

The k-means clustering algorithm is a dynamic clustering algorithm that clusters the feature set $\Phi = \{M_1, M_2 \cdots M_i \cdots M_n\}$ and uses the clustering center to form a visual dictionary $\Psi = \{\varphi_1, \varphi_2 \cdots \varphi_i \cdots \varphi_m\}$, where m is the size of the visual

* Corresponding author

dictionary and φ_i is the *i*-th visual word. Assuming the *i*-th cluster ϕ_i contains N_i sift (Lowe 2004) features, the *i*-th visual word φ_i is the mean of the sift features in ϕ_i, that is,

$$\varphi_i = \frac{1}{N_i}\sum_{\eta\in\phi_i}\eta \tag{1}$$

The sum J_e of squared errors of all the clusters is

$$J_e = \sum_{i=1}^{m}\sum_{\eta\in\phi_i}\|\eta-\varphi_i\|^2 \tag{2}$$

The sum J_e of squared errors generated when using m cluster centers $\varphi_1,\varphi_2,\cdots,\varphi_m$ to represent a set $\phi_1,\phi_2,\cdots,\phi_m$ of m sift features is measured. It is easy to see that when J_e is the smallest the clustering result is the best.

Therefore, the process of generating the visual dictionary by using the k-means clustering algorithm can be summarized as follows:

1) Initializing m cluster centers of the feature set Φ to calculate the mean of each cluster $\hat{\varphi}_1,\hat{\varphi}_2,\cdots\hat{\varphi}_m$ and the sum J_e;
2) Selecting a sift feature, which belongs to the set ϕ_i;
3) If the size N_i of the collection ϕ_i equal 1, then go to 2), otherwise continue;
4) Calculating

$$\rho_j = \begin{cases}\dfrac{N_j}{N_j+1}\|\eta-\hat{\varphi}_j\|^2 & j\neq i\\ \dfrac{N_j}{N_j-1}\|\eta-\hat{\varphi}_j\|^2 & j=i\end{cases} \tag{3}$$

5) For all $j\in[1,m]$, if $\rho_k\le\rho_j$, η in ϕ_i will be moved to ϕ_k;
6) Recalculating $\hat{\varphi}_k,\hat{\varphi}_i$ and J_e;
7) If the continuous iteration N times the J_e value unchanged, then stop, otherwise, go to 2) to continue the calculation.

Thus, a visual dictionary can be generated using k-means clustering algorithm (Kanungo et al., 2002). The above process can be optimized to enhance the descriptive power of global descriptors by vocabulary tree (Nister et al., 2006). The research shows that as the number of visual words increases, the descriptive ability of the global image descriptor will be better, and one of the goals of the vocabulary tree is to increase the number of visual words (Arandjelovic et al., 2012). The basic idea of the vocabulary tree is to generate more visual words using the hierarchical k-means algorithm. The hierarchical k-means algorithm uses local k-means algorithm to get local clustering, and then recursively k-means clustering for each class until the number of target clustering centers is obtained. Compared with ordinary k-means clustering, this clustering method can get more clustering centers and the local feature mapping of each image consumes less time because of the tree structure.

2.2 Global image descriptors based on fisher vector

Fisher vector as a global description of the image is also widely used in the image clustering and retrieval field. The global descriptor of the image based on fisher vector is encoded by fisher kernel (FK) (Liu et al., 2015). The generated vector is called fisher vector (FV) which is the global descriptor of the image. The following is a brief introduction.

(1) Fisher kernel

For the data set $X=\{x_1,x_2,\cdots,x_T\}$, the probability density function p_λ, X can be characterized as a log-likelihood gradient function

$$G_\lambda^X = \nabla_\lambda \log p_\lambda(X) \tag{4}$$

where $\lambda=[\lambda_1,\lambda_2,\cdots,\lambda_M]\in\mathbb{R}^M$ denotes M parameter vectors of p_λ. It is not hard to see, $G_\lambda^X\in\mathbb{R}^M$, that the dimension of G_λ^X depends only on the number of M parameters in the parameter set λ, but has nothing to do with the size of the data set X.

According to the information geometry theory, the parameter family of distribution $\{p_\lambda,\lambda\in\Lambda\}$ can be regarded as a Riemannian manifold M_Λ with local measure, then the local measure can be given by fisher information matrix $F_\lambda\in\mathbb{R}^{M\times M}$ as

$$F_\lambda = E_{x\sim p_\lambda}\left[G_\lambda^X G_\lambda^{X'}\right] \tag{5}$$

Defining the measure of acquisitiveness between two samples X and Y, fisher kernel is as follows

$$K_{FK}(X,Y) = G_\lambda^{X'}F_\lambda^{-1}G_\lambda^Y \tag{6}$$

Since F_λ is symmetric positive definite matrix, so there exists Cholesky decomposition $F_\lambda = L_\lambda' L_\lambda$, then fisher kernel can be expressed as the dot product of two vectors

$$K_{FK}(X,Y) = \Im_\lambda^{X'}\Im_\lambda^Y \tag{7}$$

where

$$\Im_\lambda^X = L_\lambda G_\lambda^X \tag{8}$$

Let $\Im_\lambda^X$ be the fisher vector of data set X.

(2) Global image descriptor based on fisher vector

It can be seen from the generating process of the fisher vector that the set of feature points in a single image can be used as input data set $X=\{x_1,x_2,\cdots,x_T\}$, where x_i represents the corresponding feature descriptor and T represents the number of feature points.

The probability density function p_λ can be represented by a Gaussian mixture model (GMM) (Zivkovic 2004; Ranmussen, 2000). GMM can approximate any continuous probability density distribution with arbitrary precision, so it can be used to model the image features and characterize the corresponding probability density function. The GMM parameter set with K Gaussian units can be expressed as $\lambda=\{\omega_k,\mu_k,\varepsilon_k\,|\,k=1\cdots K\}$, where $\omega_k,\mu_k,\varepsilon_k$ is the weight vector, mean vector and covariance matrix of the k-th Gaussian unit, respectively.

$$p_\lambda(x) = \sum_{k=1}^{K}\omega_k p_k(x) \tag{9}$$

where, p_k represents the k-th Gaussian unit

$$p_k(x) = \frac{1}{(2\pi)^{D/2}|\varepsilon_k|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_k)'\varepsilon_k^{-1}(x-\mu_k)\right\} \quad (10)$$

It can be seen that the estimation of GMM parameters is a crucial step in generating the image fisher vector. In order to reduce the model parameters to be estimated and reduce the computational cost, the covariance matrix of the model is assumed to be a diagonal matrix. Then, the expectation maximization (EM) algorithm can be used to estimate the parameters. Finally, the fisher vector global descriptor of the image could be calculated.

2.3 The vector of locally aggregated descriptors

The vector of locally aggregated descriptors (VLAD) is a global description method that quantifies local feature points into visual words, which same to the BoW and Fisher methods (Jégou et al., 2010). However, the VLAD is computed by adding up the difference between the local feature and its corresponding center, and contains more abundant feature distribution information. It can be understood as a compromise between BoW and Fisher methods. In the BoW method, the local features are firstly clustered by k-means, and then are represented by the cluster centers. The resulting descriptors may lose more information. In the Fisher method, using the GMM to model the local features, considering the distance from each local feature to all the cluster centers, using the linear combination of all cluster centers to represent the local features, which the generated process also results in the loss of some information. However, the VLAD method combines the advantages of both BoW and Fisher methods. It not only considers the nearest cluster center to the local feature which keeps the distance between them, but also considers each dimension value of the local feature. So, it has more detailed description of the local information of the image.

Assuming the k cluster centers $(c_1,L\ ,c_j,L\ c_k)$, which obtained by k-means clustering method, are used for the local feature points of the image, the procedure for generating the VLAD descriptor u is as follows:

1) Calculating the difference between each local feature point x_i of the image and the corresponding cluster center, and then calculating the cumulative sum u_j in the cluster.

$$u_j = \sum_{x_i \in N} (x_i - c_j) \quad (11)$$

where N is a set of feature points corresponding to the cluster center c_j .

2) Combining all vectors u_j into a single long vector, and then normalizing l_2 norm to get VLAD descriptor u .

The dimension of VLAD descriptor is $k \times n$, and n is the descriptor dimension of the local feature point. In the application, principal component analysis (PCA) can be used to reduce the dimension of the global descriptor.

2.4 Aggregation descriptors based on edge features

The edge feature points in the image contain the main semantic information of the image. The aggregation of the local feature descriptors with edge feature points can get the global descriptor which containing strong semantic information. Therefore, we firstly extract the edge information from the image to obtain the corresponding edge feature points, and then use feature descriptors to describe these edge feature points. VLAD algorithm is used to aggregate the feature descriptors so that the aggregated descriptor with edge feature points can be obtained. At the same time, considering the overall content of the image, we use the feature points of the low-resolution sampling pyramid image to generate VLAD descriptor. The two descriptors are combined to obtain a vector of edge-based locally aggregated descriptors that take full consideration of the image detail information. In practice, most images contain rich edge feature information. Without further filtering, feature descriptors extraction will consume a large amount of computation and memory. Considering the complexity of the computation process and the integrity of the detail information, feature points extraction can be performed first, and then the feature points with the edge feature are selected from these feature points to obtain the feature descriptors with the edge representation feature. First features extraction and then the edge features selection, the essence is to remove the feature points without significant physical meaning, resulting in more obvious semantic information of the global descriptor. It has the equivalence of results with the first extracting edge feature points and then generating feature descriptors.

To sum up, the generation step of the proposed algorithm can be optimized as follows:

1) Firstly, the image feature points are extracted to obtain the feature point set T_1 ;
2) Edge point detection is performed on the feature points in the set T_1 to obtain the edge point set Q_1 ;
3) VLAD algorithm is used to aggregate the set Q_1 to obtain the aggregation descriptor u_1 ;
4) Spatial pyramid down-sampling of the image is performed, and then feature extraction is performed to obtain the point set Q_2 ;
5) Aggregating the set Q_2 by using the VLAD algorithm to obtain the aggregation descriptor u_2 ;
6) The descriptors u_1 and u_2 are synthesized to obtain the final image global descriptor u .

The processing flow shown as in Figure 1.

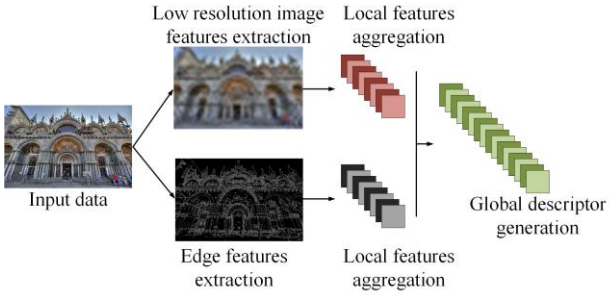

Figure 1. The proposed processing flow

3. EXPERIMENTAL RESULTS AND ANALYSIS

3.1 Global descriptors experiments

Clustering analysis between the images is essentially to find the same type of images automatically, so there are two main indexes to evaluate the clustering algorithm: first, the number of false clustering results, that is, take the error rate; second, the number of correct clustering images which lost in the clustering results, that is, the abandonment truth rate. The former index

evaluates the locality of the clustering results and mainly characterizes the in-class performance of the clustering algorithm. The latter index evaluates the integrity of the clustering results and mainly characterizes the inter-class performance of the clustering algorithm. In the experiments, we use the above two indicators as the main criterion. A comparison is made between the proposed descriptor, the traditional VLAD descriptor and the Fisher descriptor, to evaluate the quality of the global descriptor based on the edge features proposed in the paper. Clustering is performed by using the classical k-means algorithm.

The Leaves dataset from Caltech data was used for comparative experiments. The Leaves dataset contains three similar leaf images, the use of its experiments can effectively distinguish between the descriptor description of the details and the overall expression. The Leaves dataset has a total of 186 images, of which 66 A-type leaves, 60 B-type leaves and 60 C-type leaves. Some images are shown in Figure 4.

In the experiments, we use the Fisher descriptor, VLAD descriptor and the proposed descriptor to describe the Leaves dataset and then use the k-means algorithm to carry out the clustering. The Fisher descriptor is generated by training the GMM parameters in advance, using 400 Internet images downloaded from the Flickr website for parameter learning, and setting the number of clusters to 256; for the k-means parameter of the VLAD descriptor and the proposed descriptor, the same 400 images are used for parameter training, and the same number of clusters is set as 256; when the final image global descriptor k-means clustering is performed, the number of clusters is set to 3. In addition, the local descriptors used in this paper are all characterized by 128-dimension sift descriptors.

The results are shown in Figure 2 and Table 1. In Figure 2, the horizontal axis represents the cluster types under different descriptors, the vertical axis represents the accumulation of the leaves of the corresponding species in each cluster category, the different colours represent different leaf types, and the column lengths represent the number of the corresponding images. The data in Table 1 correspond to the number of specific clusters in different leaf classes in different clustering results.

Descriptors	Types	A	B	C
Fisher	I	9	10	5
	II	56	50	40
	III	1	0	15
VLAD	I	12	12	1
	II	50	46	48
	III	4	2	11
Ours	I	4	53	27
	II	60	5	17
	III	2	2	16

Table 1 Global descriptor experimental results

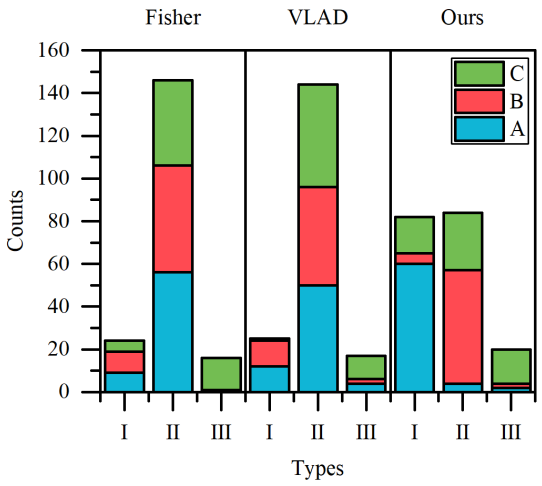

Figure 2 The situation analysis of the global descriptors experimental results

3.2 Result analysis

Ideally, the three clusters should correspond to three leaf types. As can be seen from the statistical results in Table 1 and the intuitive illustration in Figure 2, both the Fisher descriptor and the VLAD descriptor cannot cluster the Leaves dataset well, all of which contain more type A and B leaves, while most of the A, B and C leaves are assigned to class II. While the proposed descriptor is relatively well classified, that most of the leaves of type B are classified as class I and predominate, and most of the leaves of type A are assigned to class II and constitute the major component, type C leaves are assigned to class III and occupy the main ingredient.

The accuracy and error rate of each descriptor for each type of leaf clustering results are calculated by using the above results, to describe clustering results more intuitively. The leaf type defined in the cluster results, which as the main component, is the finally result representative of the cluster type. The accuracy rate (AR) of the calculation is $\eta = m / M$, where m is the number of correct clustering images in the class and M is the total number of clustering images in the class. The error rate (ER) is the average of the false positive rate (FPR) ε and the negative true rate (NTR) γ of the class. That is, taking the false positive rate as $\varepsilon = r / M$, where r is the number of correct clustering images in the class; $\gamma = w / I$ is the negative true rate, where w is the number of images that cannot be divided into the class and I is the total number of images that the class should correspond. It can be seen from the above definition that the higher the accuracy rate, the lower the error rate, and the better the overall performance of the descriptor. According to Table 1, the specific calculation results are shown in Figure 3.

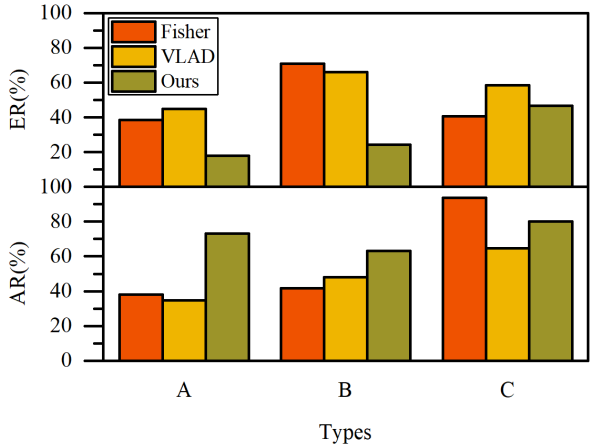

Figure 3 Global descriptors experimental AR and ER

As can be seen from Figure 3, the accuracy rate of clustering results using the proposed descriptor is higher for the leaf types A and B than for the other two descriptors, and the accuracy rate of clustering for the leaf type C is between two descriptors. However, it should be noted that the clustering results of the proposed method have the highest number of correct clustering type C leaves. The error rate of the types A and B in the proposed descriptor clustering results is much lower than the other two descriptors results. The clustering results of type C leaves are slightly higher than those of Fisher descriptor. Overall, the results of the experiments on the Leaves dataset, which can distinguish descriptor detail description ability, show that the detail description ability and the overall expressive ability of the proposed descriptor are superior to the Fisher descriptor and traditional VLAD descriptor.

4. CONCLUSIONS

In this paper, an edge-based local feature aggregation method is proposed. In view of the problem that image global descriptor cannot be well expressed in image semantics, the algorithm proposes a global aggregation descriptor algorithm based on edge feature points in combination with the specific local content and global content of the image to improve the detail expression and overall expression of the global image description. Using Leaves dataset, the experimental results show that the proposed algorithm has high accuracy and good stability, and can accomplish clustering tasks.

Figure 4 Leaves dataset diagram

ACKNOWLEDGEMENTS

The Leaves dataset were provided by the Computational Vision at Caltech (Leaves 1999): http://www.vision.caltech.edu/archive.html. The 400 Internet images dataset were provided by the Flickr: https://www.flickr.com.

REFERENCES

Arandjelovic., R., Zisserman, A., 2012. Three Things Everyone Should Know to Improve Object Retrieval. In: *IEEE Conference on Computer Vision and Pattern Recognition*, Providence, USA, pp. 2911-2918.

Csurka, G., Dance, C., Fan, L., Willamowski, J., and Bray, C., 2004. Visual categorization with bags of keypoints. In: *European Conference on Computer Vision*, Prague, Czech Republic, pp. 1-2.

Jégou, H., Douze, M., Schmid, C., and Pérez, P., 2010. Aggregating local descriptors into a compact image representation. In: *IEEE Conference on Computer Vision and Pattern Recognition*, California, USA, pp. 3304-3311.

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., and Wu, A. Y., 2002. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE transactions on pattern analysis and machine intelligence*, 24(7), pp. 881-892.

Krizhevsky, A., Sutskever, I., and Hinton, G. E., 2012. ImageNet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*, Lake Tahoe, USA, pp. 1097-1105.

Lazebnik, S., Schmid, C., and Ponce, J., 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: *IEEE Conference on Computer Vision and Pattern Recognition*, New York, USA, pp. 2169-2178.

Liu., X., Yang., S., Yang J., 2015. An Object Retrieval Method Based on Compressed Fisher Vectors. *Fire Control & Command Control*, 40(07), pp. 37-42.

Lowe, D. G., 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2), pp. 91-110.

Nister, D., Stewenius, H., 2006. Scalable Recognition with a Vocabulary Tree. In: *IEEE Conference on Computer Vision and Pattern Recognition*, New York, USA, pp. 2161-2168.

Perronnin, F., Sánchez, J., and Mensink, T., 2010. Improving the fisher kernel for large-scale image classification. In:

European Conference on Computer Vision, Crete, Greece, pp. 143-156.

Rasmussen, C. E., 2000. The infinite Gaussian mixture model. In: *Advances in neural information processing systems*, Denver, USA, pp. 554-560.

Russakovsky, O., Lin, Y., Yu, K., and Fei-Fei, L., 2012. Object-centric spatial pooling for image classification. In: *European Conference on Computer Vision*, Firenze, Italy, pp. 1-15.

Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., and Gong, Y., 2010. Locality-constrained linear coding for image classification. In: *IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, USA, pp. 3360-3367.

Yang, J., Yu, K., Gong, Y., and Huang, T., 2009. Linear spatial pyramid matching using sparse coding for image classification. In: *IEEE Conference on Computer Vision and Pattern Recognition*, Florida, USA, pp. 1794-1801.

Zivkovic, Z., 2004. Improved Adaptive Gaussian Mixture Model for Background Subtraction. In: *International Conference on Pattern Recognition*, Cambridge, UK, pp. 28-31.