An improved image classification based on K-means clustering and BoW model

Yong-Lang Liu*, Zhong Cai and Ji-Tao Zhang

Department of Computer Science, Jiangxi University of Technology, Nanchang, China
Email: aochuan@163.com
Email: caizhong@163.com
Email: zjt@163.com
*Corresponding author

Abstract: Image classification constitutes an important issue in large-scale image data process systems based on cluster. In this context, a significant number of relying BoW models and SVM methods have been proposed for image fusion systems. Some works classified these methods into Generative Mode and Discriminative Mode. Very few works deal with a classifier based on the fusion of these modes when building an image classification system. In this paper, we propose a revised algorithm based on weighted visual dictionary of K-means cluster. First, it uses SIFT and Laplace spectrum features to cluster object respectively to get local characteristics of low dimension images (sub-visual dictionary); then clusters low-dimension characteristics to get the super visual dictionaries of two features; finally, we get the visual dictionary although most of these features have been proposed for a balance role through weighting of the parent visual dictionaries. Experimental result shows that the algorithm and this model are efficient in depict image information and can provide image classification performance. It is widely used in unmanned-navigation and the machine-vision and other fields.

Keywords: image classification; visual dictionary; K-means; BoW model.


Biographical notes: Yong-Lang Liu is currently a Professor in Computer Science at Jiangxi University of Technology, Nanchang, China. His main research interests are computer networks and information management, distributed data structures and database performance.

Zhong Cai is currently a master’s student in Computer Science at education examination authority of Jiangxi province, Nanchang, China. His main research interests are computer information management, networks and database system.

Ji-Tao Zhang is currently a master’s student in Computer Science at Jiangxi University of Technology, Nanchang, China. His main research interests are computer networks and information management.

1 Introduction

Classification is a cross-research direction which involves machine learning, image processing, pattern recognition and computer vision, etc. At present, several new and efficient methods have been proposed to extract the image semantic characteristics by using BoW model (Bag-of-Words) (Chen et al., 2011; Wu et al., 2007; Yang et al., 2014; Bakshi et al., 2013; Zhu et al., 2012; Wang et al., 2016). In current image classification system, there a need to reduce the inconsistencies between different levels of features (Zhou et al., 2015; Xu et al., 2011; Lu et al., 2016; Zhu et al., 2012). Among them are BoW models and Support Vector Machines (SVM) as the main methods of image classification performance (Fei-Fei and Perno, 2005; Perronnin and Dance, 2007; Tirilley et al., 2008; Chen et al., 2013). The study of image classification methods can be divided into two categories: model-based generation (Generative Mode) and based on the discriminant model (Discriminate Mode), according to the distribution of the image feature modelling.

In the model based on the generated image classification method, the most typical way is Gaussian mixture model (GMM); Vailaya et al. (2001) apply GMM combined Bayesian classifier (Bayesian classifier) that completed the hierarchical
classification of the image. Wajeeed and Adilakshmi (2012), Zhang et al. (2012), and Zhao et al. (2010) in text classification use semi-supervised learning paradigm and compare the results obtained with the supervised learning classifier’s accuracy. Fergus et al. (2005) embedded spatial information in PLSA model to improve the disadvantages of conventional PLSA model, which lacks spatial information. Gao et al. (2013) and Wu et al. (2007) proposed a model-based approach to visual language, it becomes a visual word matrix, by assuming conditions are related between visual words. Subsequently, Tirilly et al. (2008) proposed a new vision statement image representation method based on the use of space BoW model of the relationship between visual words that constitutes appropriate ‘sentence’ and then classifies the images. Yang et al. (2013) propose an object function to describe the consistency of visual words, with category information of images and spatial context of key points, and then they adopt simulated annealing algorithm to look for a suboptimal solution, which corresponds to a visual vocabulary selected from the vocabulary tree. Verma et al. (2011) present a new oRGB-SIFT descriptor, and then integrate it with other colour SIFT features to produce the novel Colour SIFT Fusion (CSF) and the Colour Greyscale SIFT Fusion (CGSF) descriptors for image classification with special applications to biometrics.

The most widely used method is the kernel method based on discriminant model; the key is to set a distance or similarity between the images of the kernel function (Sheer et al., 2013). With the emergence of BoW model, kernel methods have been widely applied to BoW model (Csurka et al., 2004; Dhandra et al., 2011). To take advantage of the spatial structure of the image information and local features of BoW model, Lazebnik et al. (2006) obtain weighted histogram by applying to the spatial pyramid matching kernel BoW model to calculate local features of sub-image, eventually made better classification results. Kang et al. (2014), by using principal component subspace-based Gaussianity estimation, based on the idea that optimal kernel parameters lead the mapped feature space close to Gaussian distribution. Kachouri et al. (2011) propose an original kernel weighting method, by depending on the relevance of kernel training rates, to ensure better classification accuracy and significantly less computation time.

In this case, taking into account that the generated model and discriminant models have their advantages and disadvantages, some experts are dedicated to the fusion of these two models to obtain better classification performance and adaptable image classification method, which is based on the hybrid model classification (Liu et al., 2011). Holub and Perona (2005) Preetha and Suresh (2014) and Mahmoud and El Hadad (2015) combined SVM and nebula model, the cost of computing target classification by Fisher nebula model. Perronnin and Dance (2007) proposed a framework of Fisher kernel and applied it to image classification, the core framework for the use of the gradient vector signal generator to model, and used discriminative mode to determine complete image classification. Ding and Chen (2012) propose a feature fusion approach based on the joint use of spectral and spatial information provided by texture features extracted from the Grey Level Co-Occurrence Matrix (GLCM). Ai et al. (2013) and Zhong and Peng (2016) propose multi-agent colour image classification architecture (MACICA). Agents within a multi-agent system (MAS) architecture are programmed to deliver specific image classification capabilities.

2 K-means clustering

The cluster analysis based on the local characteristics of image is used to find the visual dictionary (Lihe and Chen, 2015); the cluster centre was the visual words. The all words form an image’s visual dictionary, which will affect the follow-up image classification effect, so it depends on the clustering method largely that the image classification.

Generally speaking, the clustering algorithms (Zhao et al., 2012) need to meet more similar in the same clustering objects, and less small in the different clustering objects. We could assign Q to the K clusters to minimise the sum of squares in the K-means clustering algorithm. The mathematical expression is as follows (1):

$$
\min \sum_{i=1}^{K} \sum_{x_j \in C_i} \text{dist}(C_i, x_j)^2
$$

in which K is the number of clustering centre, C_i is the centre of clustering, i = 1, …, K, X_j is the object of clustering, j = 1, …, Q.

3 SIFT features

One way to bring these closer together is to use spot detection feature description, feature generation and feature matching, our descriptor is through extract the image local invariant features to create SIFT feature vectors which have good invariance in scale, rotation, illumination and perspective transformation.

Figure 1 shows that process used here; in next section, the implementation of the solution will be introduced in detail.

1 Constructing scale space and detecting extreme

To make the feature points satisfy the scale invariant of the change those feature points in SIFT detecting algorithm. Lowe optimised for its use by LoG, and comes up an approximate algorithm DoG. The algorithm is to find the point that than its adjacent 26 points bigger or smaller, we selected the point as character point. In the next steps, we need to filter the candidate and remove those points unstably.

2 Precise positioning the feature points

There are a large number of unstable characteristic points in constructing scale space; we can get feature points of precise positioning through the elimination and selection of stable point.
An improved image classification

Figure 1  The flow chart of SIFT algorithm

Constructing scale space and detecting extrema

Precise positioning the feature points

Determine the cardinal direction of feature

Generate SIFT feature descriptor

The Taylor expansion, namely that $D(x, y, \sigma)$, is the candidate point, as follows (2).

$$D(x) = D + \frac{\partial D}{\partial X} X + \frac{1}{2} X^T \frac{\partial^2 D}{\partial X^2} X$$

(2)

In the formula $X(x, y, \sigma)$ is the offset of the sample point. We take the derivative and make it equal to zero, it is concluded information that the position of the extreme point.

$$\hat{X} = -\frac{\partial^2 D}{\partial X^2} \frac{\partial D}{\partial X} X$$

(3)

Formula (3) into the formula (2) and keep only the first two terms:

$$D(\hat{X}) = D + \frac{1}{2} \frac{\partial^2 D}{\partial X^2} \hat{X}$$

(4)

when $D(\hat{X})$ is less than a certain threshold, which will be seen as lower contrast point and to be eliminated, in Lowe’s experiments the threshold value is 0.03.

3  Determine the cardinal direction of feature

For each direction determine the feature points of distribution to ensure that SIFT descriptor has the image rotation invariant. For convenience, we will scale space image $L(x, y, \sigma)$ as $L(x, y)$, so calculate the gradient direction of $m(x, y)$ and $\theta(x, y)$ in point $(x, y)$ as following:

$$m(x, y) = \sqrt{\left(L(x+1, y) - L(x-1, y)\right)^2 + \left(L(x, y+1) - L(x, y-1)\right)^2}$$

$$\theta(x, y) = \arctan \frac{L(x, y+1) - L(x, y-1)}{L(x+1, y) - L(x-1, y)}$$

(5)

Experiment in the centre on point $(x, y)$, with a radius of $1.5\sigma$ to calculation gradient histogram and statistics, the range is $0°$–$360°$, and every ten degrees can be divided into a column (bin), together can take 36 bins. The nearer the feature points of neighbourhood points, the larger the gradient of its contribution to the greater weight, whereas the smaller contribution. In this the peak of gradient direction histogram was defined as the principal direction of the feature points.

4  Generate SIFT feature descriptor

First of all, we will coordinate plane rotary to coordinate the principal direction, then choosing a window with a $16 \times 16$ be centred with feature points and divided into $4 \times 4$ sub-areas. Then we come to calculate the direction of an eight gradient histograms in each area. Finally, we group and order in each region combinations, we can get a feature vector of $4 \times 4 \times 8 = 128$-dimensional. Figure 2 shows the diagram of characterised descriptor.

Next, we need to normalise the feature vector. Setting $D = (d_1, d_2, d_3, \ldots, d_{128})$, after normalisation processing it can be:

$$\bar{D} = \frac{D}{\sqrt{\sum_{i=1}^{128} d_i^2}} = (\bar{d}_1, \bar{d}_2, \bar{d}_3, \ldots, \bar{d}_{128})$$

(6)

After the above four steps, we got the 128-dimensional feature descriptor of SIFT.

4 Visual dictionary construction and characteristics of quantification

The BoW model is originated the text of the document retrieval, it classifies and identifies document by simply word frequency statistics document. As a document is made up of words, an image can be seen as a series of the collection of local characteristics, and in order to have the characteristics of high dimension the real value of the local features into a discrete visual words, you need to build the image local characteristics of the visual dictionary, thereby to quantify the local characteristics in image space, as described in the following steps.

1  Extracted from typical image dense local characteristics, build local characteristics of the image library of large corpora.

2  Local feature extraction feature space will be quantified for the discrete nature of clusters; the centre of the
cluster is visual words, and visual dictionary that is constituted by all of the visual word.

3 For a given image, to calculate the distance of the local features and visual words, and the statistical characteristics of each local nearest visual word frequency, so according to the frequency statistics can be a partial feature contains a large number of high dimensional data of the image is converted into a number of visual word list.

4 A visual words appear frequency of the word bag of an image histogram can be said. However, it may exist in the traditional visual words larger quantisation error, here refers to a local characteristics and the difference between the assigned visual words.

5 Improved $K$-means clustering and the generation method of visual dictionary

Firstly, hierarchical clustering algorithm is used to clustering kinds of image, obtained part of the visual dictionary of each category by kinds of image. To make each type of image to keep in balance, we will keep consistent in each type of image clustering; next, we could get parent visual dictionary by clustered using $K$-means algorithm again. In this paper, we get two parent visual dictionaries for the same image because we use two kinds of image features to structure the visual dictionary. By using $K$-means clustering analysis method, classifies each of training image respectively, we get part of visual dictionary $C_{iL}$ and $C_{iSK}$. $C_{iL}$ is the centre of Laplace spectrum structure as the $i$th element of training image, $C_{iSK}$ is the centre of SIFT local feature as $i$th element of training image. Here $K_i$ is the number of $i$th element of training image, and $i = 1, 2, ..., M$, here $M$ is the number of category of training image.

$$C_{iL} = \text{Cluster}\{C_{iL}, C_{iL}, ..., C_{iL}, ..., C_{iL_u}\}$$
$$C_{iSK} = \text{Cluster}\{C_{iSK}, C_{iSK}, ..., C_{iSK}, ..., C_{iSK_u}\}$$
$$C = \partial C_{iSK} + (1 - \partial)C_{iL}$$

Here C is the total visual dictionary, $\partial$ expresses the weight of clustering, and here $\partial = 0.8$. Based on the above $C_{iSK}$ and $C_{iL}$, each image by VQ algorithm could expressed as a number of visual words, we get two histograms of complementary vectors by each image histogram statistics, there are VL and VS, so we can train and classify through construct classifier.

6 Experiments

In order to verify the feasibility and the accuracy of image classification, the Laplace spectrum structure feature based on the model of BoW, we selected from ALOI image library (http://aloi.science.uva.hi/) seven kinds of real image sequences respectively windmill, house, tower, inn and frame. As shown in Figure 3, each type of images was selected and randomly selected from 10 to 60 for training, 50 for testing. According to our classification process we extract SIFT local features and Laplace structural characteristic for each image, to get the final image combining BoW model, which is represented as a vector image, and then transporting it to LIBSVM classifier for image classification (BoW + SIFT + Laplace spectrum), the value averaging operation after ten experiments and test comparison with the method of BoW + SIFT and adjacency spectrum + NMF. As Figure 4 shows, the experimental results indicated that the method used in BoW + SIFT + Laplace spectrum classification has a high accurate rate, among them, those black boxes indicate the accuracy of the classification of each type of image. The results are shown in Table 1.

Figure 3 The image of testing
7 Simulation analysis

As can be seen from Figure 4, the accuracy of classification is much better than other methods in which colours of the background are not complicated. This movi-image, frame, windmill, reaching more than 98%; images inn, house1, house2 classification accuracy is low, only 94%. All images have lower classification accuracy for other images; results indicate that the proposed algorithm has better classification performance for foreground or background simply image.

Similarly, compared to the other two kinds of classification methods, it can be seen from Table 1, the images inn and house1, house2 there are high similarity: for tower, frame and windmill images similarity crosstalk appeared in the process of classification, so image classification accuracy, NMF algorithm based on the average number of errors amounted to 88, and the number of errors BOW + SIFT is 34, which is much higher than BOW + SIFT + Laplace algorithm, as can be seen from Table 1, the average number of errors is only 15 of this method, because this method has described image features and can detail the structure of local image features, so it is better to eliminate inter-image similar to a class of interference. So it gets better classification results.

The proposed method classification accurately was about 96.5% of the time, higher than the other two methods on average classification accuracy, indicating this algorithm depends only on SIFT or Laplace spectral characteristics that cannot be well-described structural features of images. So their classification is not ideal.

8 Conclusions

Work of this paper is based on the extracted uniform division Laplace spectral characteristics, and pointed out lacks K-means clustering methods and make improvements for building visual dictionary, we put forward a method of constructing visual dictionary based on a weighted hierarchical K-means clustering strategy. It uses SIFT and Laplace spectrum features to cluster the image hierarchically to get part of visual dictionary of each category of images; then to cluster all part of visual dictionaries for two super features; then, we used weight distribution between two super visual dictionaries, then to get the final visual dictionary. Finally, we complete the image classification experiments combined LIBSVM classifier, and make analysis of the experimental results.

However, with the development of modern networks and the explosive growth of life, the content of complex digital images, combined BoW image classification task, there are still many issues that need attention and resolution. To propose a better local feature aggregation method is we are going to study the characterised state of BoW model image is disordered, but in the extracted local features, there is some local characteristics information richer than other local features; therefore, in the image feature convergence strategy, at first we can partial image feature based on many kinds of priority order, then get the image feature of the chain representation, we believe that the image obtained in this way is more summarised image semantic content and information.

References


