Milk Somatic Cells Recognition based on Gray-Scale Difference Statistics

Xiaoli Zhang [1,*], *Heru* Xue[1], *and Xiaojing* Gao[1]

[1]College of Computer and Information Engineering, Inner Mongolia Agricultural University, China

Abstract. In order to solve the problems in recognition of milk somatic cells and improve the recognition efficiency and accuracy, we put forward a classification algorithm based on the gray-Scale difference Statistics(GLDS). First, the algorithm extracts the gray difference matrix of the image, and from the matrix we can calculate the texture features of cell images. Then uses the K-means algorithm to segment the cells, extracts morphological features from the nucleus. We extract the 8 morphological feature parameters, 4 texture feature parameters, and use the K-fold cross-validation and random forest classification(RF) to classify samples. Experimental results show that the correct rate is 95.36% when using the random forest classifier.

1 Introduction

Through scientific research, it is shown that the type and quantity of milk somatic cells is an important indicator for milk quality assessment and diagnosis of mastitis. Somatic cells in milk are mainly derived from blood leukocytes and can be divided into lymphocytes, neutrophils and macrophages, and the other small part is the epithelial cells that are shed from the breast tissue, which is very small, accounting for 0~7% of the total number of cells[1]. Although the classification of somatic cells can be achieved by artificial methods, Although the classification of somatic cells can be achieved by artificial methods, there are several problems with this method: 1. Inefficiency; 2. Subjective factors of people may lead to inaccurate judgment; 3.Visual fatigue leads to wrong judgments and so on.

With the development of image processing technology and pattern recognition technology, it is feasible to realize computer automatic recognition of the Somatic microscopic image, and have a higher recognition rate. This paper proposes a recognition classification algorithm that combines the regional characteristics of cell images with the statistical characteristics of gray-level difference. First, preprocessing the original cell image, and calculate four texture features by using gray-scale difference statistics. Then use k-means cluster segmentation algorithm to divide that cell image into three parts, and extract morphological features from the divided nucleus; Classification recognition through K - fold cross validation and random forest algorithm.

2 Gray-scale different statistics

Gray-scale difference statistics are used to describe the change in gray levels between each pixel and its neighboring pixels in the texture image. The probability of appearance of gray-scale differences within a certain range of neighborhoods, can reflect the degree of association between different pixels. This is the basic principle of the gray difference statistical algorithm[2-3]. Suppose A: (x, y) is a point in the image, B: (x+Δx, y+Δy), there is only a small distance between A and B. The gray-scale different value of two points is[4]:

$$g\Delta(x,y) = g(x,y) - g(x+\Delta x, y+\Delta y) \quad (1)$$

gΔ(x, y) is gray-scale different, Suppose all possible values of the gray-scale difference are m. Let the point (x, y) move on the image, calculate the number of times that gΔ(x, y) takes each value. And then calculate the histogram of gΔ(x, y), The probability p(i) of gΔ(x, y) can be known from the histogram. If the i-value is smaller, p(i) is larger, then the texture is rough. On the contrary,if the probability is more balanced, the texture is fine.

Based on the gray difference histogram, some feature parameters can be calculated, these parameters quantify the characteristics of the image. The main characteristic parameters are: Contrast(CON), Angular second-order moment(ASM), Mean(MEAN), and Entropy(ENT).

(1) Contrast(CON):

$$\mathrm{CON} = \sum_{i=0}^{255} i^2 p(i) \quad (2)$$

CON represents the depth of image clarity and texture grooves. Texture grooves deep, the value of the contrast is greater; The grooves shallow, the value of contrast is smaller.

(2) Angular second-order moment(ASM):

$$\mathrm{ASM} = \sum_{i=0}^{255} p^2(i) \quad (3)$$

* Corresponding author: 727168779@qq.com

ASM represents the uniformity of gray-scale distribution of the image, is the sum of the squares of the gray difference frequency.

(3) Mean(MEAN):

$$\text{MEAN} = \frac{1}{m}\sum_{i=0}^{255} ip(i) \quad (4)$$

MEAN represents the overall gray-scale value of the image. If the image has more luminance regions, the MEAN is higher.

(4) Entropy(ENT):

$$\text{ENT} = -\sum_{i=0}^{255} p(i) \quad (5)$$

ENT is a measure of randomness. it is a measure of the amount of information in an image. When the gray difference histogram distribution is more dispersed, the ENT is larger.

In the formula above, when p(i) is flat, ASM is small and ENT is large. If p(i) is distributed near the origin, MEAN is smaller.

3 Regional feature extraction

Features are quantitative descriptions of cells. In the process of identifying cells, it is the most important to select properly features. for example: the smaller feature difference of the same types of cells is better, it is the opposite for different types of cells[5].

3.1 Segmentation of Milk Somatic Cells

The nucleus of milk somatic cells contain most of the features, therefore, the cells and nucleus must be separated from the background. This paper uses the K-means image clustering segmentation algorithm. K-means is the partition-based clustering method[6]. The idea of the algorithm is to cluster the K points in space, and the objects near the center point are classified. Through the method of iterative, update the values of each cluster center step by step, until the best clustering results.

3.2 Regional Feature

The image regional feature contain many important information. These four types of cells' nucleus have large differences in the size and shape. The nucleus of lymphocytes is large and round; The nucleus of neutrophils is lobulated, dumbbell-shaped or elliptical; The nuclear characteristics of macrophages and epithelial cells are similar, they are all round or oval. This paper extracts features from the following aspects:

(1) Area:

Cell area is usually extracted in two ways. One is to calculate the extracted target pixels, this method is simple and accurate. The second is the use of a chain code table, which considers the target object to be composed of a number of closely adjacent horizontal segments. In this way, the sum of the length of all horizontal segments is the area. The formula is:

$$\text{S} = \sum_{k=1}^{m}(X_{k2} - X_{k1}) \quad (6)$$

In the formula: S is the area of cell; X1, X2 are the abscissas of the two endpoints of the line segment; k is the marked variable of the line segment; m is the number of segments.

(2) Perimeter:

The perimeter of the cell is represented by the perimeter of the boundary contour of the contained region. The formula is:

$$\text{P} = N_l + \sqrt{2}N_h \quad (7)$$

In the formula: P is the perimeter of the nucleus; N is the number of lines that make up the line between the horizontal and vertical coordinates on the nucleus boundary; N is the number of line segments formed by two adjacent points diagonally on the cell boundary.

(3) Squareness:

$$\text{R} = \frac{S_H}{L*W} \quad (8)$$

In the formula: SH is the area of the nucleus; L, W is the length and width of the smallest circumscribed rectangle.

(4) Elongation:

Elongation indicates whether the nucleus is close to a circle. If the value is 1, then the nucleus is round; If the value is smaller, the nucleus area is slender.

$$\text{E} = \frac{W}{L} \quad (9)$$

(5) Circularity:

Circularity indicates the complexity of the nucleus. When the area is constant, the smaller the perimeter, the closer the cell nucleus is to the circle; On the contrary, the larger the perimeter,the more complicated the shape.

$$\text{C} = \frac{p^2}{4\pi S} \quad (10)$$

(6) Nucleus- to - cytoplasm ratio

There is a big difference in the nucleus-to-cytoplasm ratio of different types of cells. lymphocytes have large nucleus and Less cytoplasm, the ratio is closer to 1; Macrophages have smaller nucleus and more cytoplasm, so the smaller the ratio of macrophages.

4 Classification and identification

4.1 Random forest algorithm(RF)

The Random Forest was first proposed by Leo Breiman and Adele Cutler. The algorithm combines Breiman's' Bootstrap aggregating' idea with Ho's 'random subspace method'. In machine learning. Random forest is a classifier that uses multiple decision trees to train and

predict samples. These decision trees are independent of each other, the growth of trees and the selection of training samples are both using random methods, and reduces the higher variance of the tree structure classifier[7]; Each decision tree is not related to each other and is independent of each other. When there is a new sample input, each tree in the random forest determines the sample, then all classifiers follow the majority rule, and determine the classification result. Random forests are highly efficient, insensitive to noise, and many other excellent inherent properties and ideal classification effects. Make it use frequently in machine learning, and the effect is remarkable[8].

4.2 K-fold Cross Validation

Cross validation is a common method for measuring accuracy of models[9]. It is also a statistical analysis method used to verify the performance of the classifier. When the available data is relatively small, through effective reuse of data to determine the stability of the model. The cross-validation idea[10] is to group raw data, a portion as a training set, another as a test set, training Classifiers with training Sets, and using the test set to test the training model. This method is used to evaluate the performance of classifier.

K-fold cross-validation is one of the common forms of cross-validation[11]. The algorithm divides the data sets into K parts on average, each time take a sample from the sample set as a test set, other as training set; Repeat K times,and take all subsets in turn for the test set. Experimental results take the average of all data.

4.3 Experimentation Design

The experimental data includes 120 groups of samples, divided into four types of cells, and 30 samples of each type of cell. In order to improve the evaluation criteria and increase the reliability of the experiment, this paper uses two experimental programs:

Solution A: Compare the recognition effect of GLDS and gray level co-occurrence matrix(GLCM) features. Randomly select 60 training sets and 60 test sets, each experiment uses different features. After a large number of experimental tests, comparing the recognition effect of different features.

Solution B: Verify the stability of the random forest classifier. The experiment uses 5-fold cross validation to train and test the sample set. Repeat 100 experiments, and the results are averaged. Exclude factors such as instability of single experiment results. According to the test results, the average accuracy rate (ACC) and the standard deviation (STD) of cross-validation accuracy were calculated to judge the quality of the classifier.

5 Experimental results and analysis

As you can see from the table 1, The recognition rate of GLCM is 76.8%, and the recognition rate of GLDM is 88.94%, it's about 12% higher than GLCM. So as you can see, in this paper, the statistical characteristic parameters of the GLDM are effective and feasible. At the same time, after combining the features of the image region, the recognition rate is as high as 95.69%.

Tab. 1. Recognition rates under different features

Feature	Recognition Rate(%)
GLCM	76.8
GLDM	88.94
GLCM+Morphological	90.8
GLDS+Morphological	95.69

Table 2 is the recognition rate under different classifiers. It can be seen that the Bayes classifier and the K nearest neighbor (KNN) classifier have almost the same recognition rate. However, the random forest has a higher recognition rate than other classifiers.

Tab. 2. Recognition rates under different classifier

Number of amples	KNN_ACC (%)	Bayes_ACC (%)	RF_ACC (%)
120	0.91388	0.91389	95.69

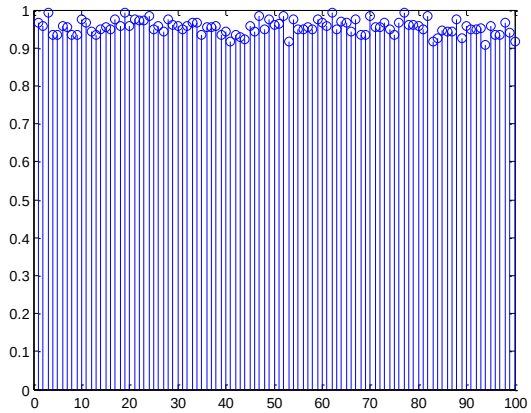

Fig. 1. 100 cross-validation recognition rates

In order to verify the stability of the random forest classifier, This experiment has done 100 times of 5-cross-validation. Figure 1 shows the recognition rate of 100 times. The recognition rate is higher than 91%, and the average recognition rate is 95.36%. Therefore, this method has a better recognition effect.

6 Conclusion

In the classification of Milk somatic cells, the traditional classification method is based on the experience and knowledge of cell testers. There is more subjectivity, and misdiagnosis occurs during the diagnosis of mastitis. This paper uses random forest and K-fold cross-validation for the identification and classification of somatic cells to determine whether it is mastitis. From the experimental

results we can see: For the extraction of textures and morphological features from somatic cells, the correct rate of identification using random forest classifiers reached 95.36%. The algorithm has higher accuracy and stronger reliability.

Acknowledgment

This work is supported by national natural science foundation of China (No.61461041).

References

1. H.R. Xue, Studies on Segmentation Methods of Milk Somatic Cell Color Images.J. IMAU (2007)
2. H. Liu, Methods Research of Flame Image Multiple Features Extraction for BOF Steelmaking Blowing Data Prediction. KUST (2012)
3. H. Liu, Y.S. Zhang., Y.H. Zhang, Texture Feature Extraction of Flame Image Based on Gray-Scale Difference Statistics. Control Engineering of China, **20**(2): 213-218 (2013)
4. X.H. Sun, *Digital image processing principles and algorithms*. China Machine Press (2010)
5. G. GARCIA, J. BERNUSSOU, Pole Assignment for uncertain systems in a specified disk by state feedback. IEEE T AUTOMAT CONTR, **40**(1): 184-190 (1995)
6. X.H. Han, Y. Hu, Research of K-means Algorithm.J. TYUT, **40**(3): 236-239 (2009)
7. X.F. Gu, Research and Application of Visual Tracking Algorithm Based on Random Forest. J. NJUST (2013)
8. W. Sun, Fine-needle aspiration diagnosis of breast neoplasms based on random forests. Journal of Computer Applications, **35**(S2): 143-145, 169 (2015)
9. Q.H. Wang, J.W. Liu, L.L. Zhang, Study on the classification of K-Nearest neighbor algorithm.J. XATU, **35**(2)：120-124 (2015)
10. J.X. Hu, G.J. Zhang, k-Fold Cross-Validation Based Selected Ensemble Classification Algorithm. Bulletin of Science and Technology, **30**(5): 924 (2013)
11. H. Xu, Y.T. Wang, D.F. Chen,*Modern Communication Network Technology*. TSINGHUA UNIVERSITY PRESS(2004)