Scholar Publishing Group
Machine Learning Theory and Practice
https://doi.org/10.38007/ML.2022.030202
ISSN 2790-0983 Vol. 3, Issue 2: 12-23

A Fine Wool and Cashmere Identification System Incorporating Decision Tree Algorithms

Yaning Yang*
Guangzhou College of SCUT, Guangzhou, China
**corresponding author*

Keywords: Decision Tree Algorithm, Fine Wool, Cashmere, Image Classification

Abstract: In practice, the identification of cashmere and wool fibers is still manual. With the expertise of experts, the shape of the fiber surface is observed and distinguished. This manual identification method is difficult because it requires at least years of training for qualified inspectors to observe a large number of fiber images, which is time-consuming and cannot avoid subjective human intervention. Although researchers have proposed many alternative identification methods, none of them can combine efficiency and cost. Therefore, it is particularly important to find a convenient and efficient identification method. The main purpose of this paper is to fuse the decision tree algorithm to study the identification system of fine wool and cashmere. This paper systematically introduces the method of cashmere and wool fiber identification based on image processing technology and feature extraction technology through fiber apparent morphological features. The process of fiber recognition is improved, and the features used for fiber recognition are obtained. The relevant theories and implementation methods are introduced in detail. Two key steps, image processing and feature extraction, are realized based on the theory. Finally, the effectiveness and feasibility of the method are verified through experimental tests, and good results are achieved.

1. Introduction

In recent years, due to the deterioration of the ecological environment and changes in goat genes, the structure of cashmere has become uneven, making it very difficult to identify cashmere and shepherd dogs. The research and identification of cashmere and other specific animal fibers are based on the classification of production areas and the establishment of cashmere fiber maps for each production area. The collection, storage, exchange and management of cashmere fiber scale characteristics in different regions are realized, which makes the identification and identification of cashmere fiber more necessary than in books, and can quickly and timely update the morphological characteristics of the fiber [1-2].

In relevant research, Kazim et al. focused on designing an objective, simple, fast, time-consuming and cost-effective method to separate wool fibers from mohair fibers by using texture analysis based identification methods [3]. Therefore, the microscopic images of wool and mohair fibers are preprocessed into texture images. The feature extraction process based on local binary pattern and deep learning are respectively used to obtain certain information from fibers. In order to identify samples, the classification based method is completed. The experimental results show that through the use of depth learning and machine learning, accurate texture analysis of this animal fiber can identify wool and mohair fibers with 99.8% and 90.25% accuracy respectively. Stefano et al. mainly studied the computer automatic measurement of cashmere diameter and the measurement method of cashmere fiber. The simulation experiment was carried out under the simulation environment of Matlab7.0, and the processing scheme suitable for the fiber image was determined. Finally, according to the preprocessed image, the automatic measurement of cashmere diameter was realized by using the subsection measurement algorithm [4]. The experimental results show that the measurement method adopted in this paper improves the measurement accuracy of cashmere diameter to a certain extent. The model features of fiber scales are mainly based on the identification of animal fibers. This paper studies six features of cashmere and wool fibers, namely, diameter, height ratio diameter, scale height, scale projection width, right angle scale thickness and scale diameter difference. Finally, Bayesian model is used to identify them. The results show that the method is very effective for the identification of ordinary wool and cashmere, as well as ordinary wool and elastic wool.

In this paper, two feature extraction methods are mainly used to identify fibers. The two feature extraction methods are introduced in detail and compared experimentally. The result shows that the fiber recognition rate based on SURF features is better, because the parameter information of fiber cannot be completely and accurately obtained during image processing, which also proves the effectiveness of SURF in cashmere and wool fiber image classification.

2. Design Research

2.1. Existing Problems

At present, the identification method of cashmere and wool used in practice is mainly based on the manual identification method of optical microscope. The principle of this method is based on the fiber appearance [5-6]. Due to the shortcomings of manual detection, such as time-consuming, labor-intensive and poor repeatability, people began to explore automatic fiber identification methods and did a lot of research work [7-8]. Among them, the automatic recognition method based on fiber microscope image is a research hotspot in recent years, but from the perspective of existing research work, there are still some deficiencies:

1) In the current research based on fiber image, the extracted fiber appearance morphological features mainly include fiber diameter, fiber surface scale height, density, shape and other measurement features. However, due to the low magnification and small depth of field of the optical microscope, and the relatively clear fiber image can only be observed under the focal plane, the image will be deformed and blurred if it deviates from the focal plane, These conditions greatly affect the accurate measurement of various parameters of fiber scales [9-10].

2) At present, the sample size used in fiber image based research is relatively small, while the dispersion of cashmere and wool fiber diameter and other characteristics is relatively large, and the characteristics of cashmere fibers from different varieties and places of origin are also different. Whether many research results are also applicable to large samples needs further verification.

3) At present, the research on the identification of cashmere and wool mainly focuses on the fiber classification into cashmere and wool. However, in practice, it is necessary to identify

different types of cashmere (such as blue cashmere and white cashmere). However, there are few literatures about these studies [11-12].

2.2. Decision Tree Algorithm

The decision tree is a prediction model for statistical analysis of data in the form of a tree. It realizes the correspondence between attributes and values among objects, and judges whether the conditions are met according to attribute information at nodes [13-14].

Algorithm execution steps:

1) Training stage: randomly select sample data and characteristic parameters from the data set in a certain proportion to form a training data set;

The selection of root node - entropy S (indicates the degree of confusion. The greater the entropy value, the worse the classification effect);

The entropy S is calculated as follows:

$$S = \sum(-p \log p) \tag{1}$$

Where p is the probability of classification results.

The information gain Δ S can be expressed by the difference in the degree of confusion between itself and the eigenvalue, and its calculation formula is as follows:

$$\Delta S = S_o - S_l \tag{2}$$

The self entropy So is based on the classification result of the tag, and the eigenvalue entropy Sl is based on each eigenvalue.

If there are many feature attributes and few corresponding samples, the information gain may be very large. Therefore, the information gain rate Sr is used to divide the categories. The calculation formula is as follows:

$$S_r = \Delta S / S_o \tag{3}$$

Non leaf nodes are selected according to the information gain of entropy value.

2) Classification: classify the decision tree by taking the root node as the starting point, and then divide it down in order until the classification results of each branch below the leaf node are the same. No further classification is required [15-16];

3) Decision tree pruning: too many branches will result in over fitting, so it is necessary to prune it. The process of pruning includes pre pruning and post pruning. Pre pruning terminates in advance when a certain condition is reached during the operation, and its operation mode is to limit the depth of the decision tree or the sample size. The post pruning starts after the program constructs the decision tree. The pruning is based on the calculation of the evaluation function value of different nodes. The evaluation function value is the product of the weight value of all node samples and the corresponding node function value. The smaller the value of the evaluation function, the better the classification effect will be by subtracting branches [17-18].

3. Experimental Study

3.1. Overview of the Development of Image Classification Technology

The development of image classification technology can be divided into three stages: (1) traditional image classification methods; (2) Modern mainstream image classification methods; (3)

The end-to-end multi-layer feature learning method is shown in Figure 1.

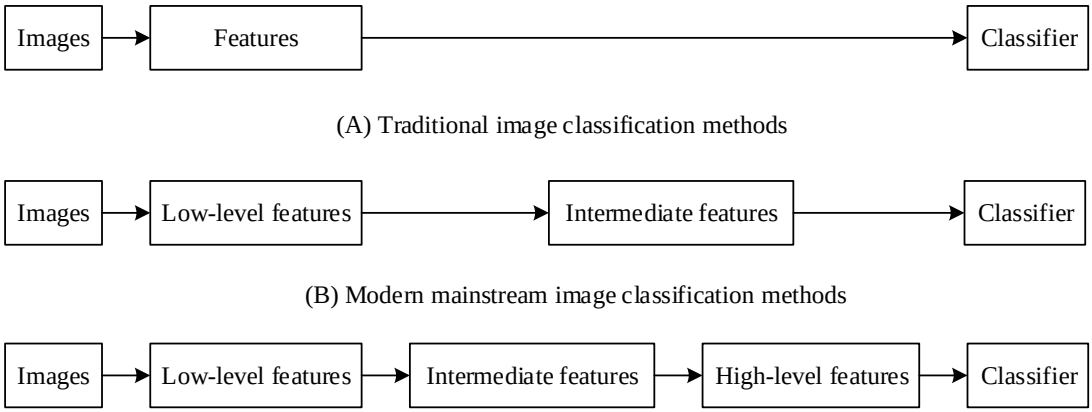

Figure 1. Three stages in the development of image classification techniques

1) Traditional image classification methods use the method of manually specifying features, as shown in Figure 2-1 (a). The process of image classification is divided into two steps: (a) manually designing feature extraction methods; (b) The classifier is used to train and test the extracted feature information.

2) The modern mainstream image classification technology divides the classification process into three steps, as shown in Figure 2-1 (b). This method also uses artificially designed features. In the first step, the designated feature extraction method is used to extract features from the image, which is called low-level features here; The second step is to express low-level features as intermediate features, usually using unsupervised learning methods (such as clustering), and then use intermediate features to describe the image; The third step is to use intermediate features to train the classifier, and then complete the image classification task. The word bag model belongs to this type of method.

3) The latest image classification method uses an end to end multi-level feature learning method. This method does not require manual design features, but starts from training data, through multi-level learning, to form a hierarchical representation of image features. Many researches have pointed out that the deeper level convolutional neural network can better obtain the features in the data, and thus obtain higher classification performance.

3.2. Experiments

1) Experimental data

The samples are cashmere and wool fiber images provided by the same cashmere group. The visual vocabulary (visual dictionary) of cashmere and wool is obtained according to the process described above. By using supervised learning training, this paper also conducts a comparative experiment on the selection of training set size, because if the proportion of test set in the total sample size is too small, the training accuracy may not be reached. If the proportion of test set in the total sample size is too large, the over fitting phenomenon may occur, which makes the final recognition effect worse. Therefore, the ratio of training set and test set is finally determined to be 7:3. Finally, more than 1100 images of cashmere and wool training set and more than 400 images of test set are selected.

2) Determination of K value

The size of the K value has a certain impact on the recognition rate. If the K value is too small, it

is difficult to effectively describe the image, and if the K value is too large, it is easy to produce an over fitting phenomenon. Generally, the selection of the K value should be less than the number of feature points of all images. According to the analysis and comparison of many experiments, the final selection of K value of 500, that is, 500 visual words, is used to describe the characteristics of cashmere and wool fiber image

3) Determination of Kernel Function

The SVM training function in the support vector machine function toolbox in MATLAB is used for training; when the kernel function is different, the recognition rate will be different.

4) Determination of L value

When extracting visual words from images, spatial pyramid method is added. Different pyramid levels will lead to different final recognition rates. Therefore, the size of L value needs to be discussed and analyzed.

5) The specific steps of fiber diameter measurement are as follows:

(1) After the central axis of the fiber is obtained, any point Pi (xi, yi) on the central axis can be obtained, i=0.12.3...., n. Search the points before and after the point at any point, and the slope of the point can be obtained according to these points;

(2) Using the slope of the point as the vertical line of the point, the intersection point with the fiber edge line can be obtained;

(3) Set the intersection point of the vertical line and the fiber edge line as C (x1, y1) D (x2, y2). At this point, the number of pixels between CDs can be obtained, that is, the number of pixels between fiber edge lines. The fiber image is 640x480. It can be seen that the original fiber image has a 100um ruler. According to the length of 100um, the number of pixels within 100um can be calculated, so that the size of a pixel can be known. Through calculation, it can be known that the size of a pixel is 0.376um. Then calculate the number of pixels between the two edge lines of the fiber, and the product of the pixel size is the obtained fiber diameter.

6) Computing and Testing Classification Functions

The classification function mainly has two parameters: support vector and constant term. The support vector is composed of data points one by one. The eigenvalues and support vectors of each picture obtained above are used for operation. The result is added to the value of the constant term to obtain the final function value of the classification function. If the value is greater than 0, it is judged as wool, and if the value is less than 0, it is judged as cashmere.

The complexity of the classification function is mainly reflected in the size of the support vector. The factors that affect the complexity are the type of kernel function and the implementation method. At present, there is no strict theoretical basis for the value taking principles and methods of variables such as kernel function type and implementation method. The control variable method is adopted, that is, the method of fixing other variables and changing the value of a variable. Many simulation tests are carried out to find the best value of each variable.

4. Experiment Analysis

4.1. Experimental Results

1) Function comparison

In this paper, linear kernel function, radial basis function (RBF) and histogram cross kernel function are used for approximate recognition and classification of cashmere and wool, and the recognition rate is shown in Table 1 below:

Table 1. Recognition rate results for cashmere wool fibres

Methods	Wordpack model			Join the Spatial Pyramid		
SVM kernel functions	Linear	Radial basis	Histogram crossover	Linear	Radial basis	Histogram cross
1	75.3%	74.2%	76.1%	83.3%	82.2%	84.1%
2	70.6%	72.4%	77.5%	84.2%	84.8%	88.4%
3	79.7%	80.2%	80.8%	83.2%	83.2%	85.8%
4	79.3%	75.2%	80.1%	83.2%	80.2%	84.2%
5	74.4%	76.7%	78.6%	82.3%	83.2%	87.2%
6	74.3%	73.8%	76.9%	80.2%	82.0%	84.6%
7	75.9%	75.5%	76.4%	83.2%	81.7%	86.2%
8	74.6%	75.2%	78.1%	80.4%	83.6%	84.2%
9	77.8%	75.2%	77.7%	79.8%	80.6%	85.2%
10	75.5%	74.2%	81.3%	81.2%	81.6%	84.2%
Average	75.64%	75.26%	78.45%	82.10%	82.29%	85.87%

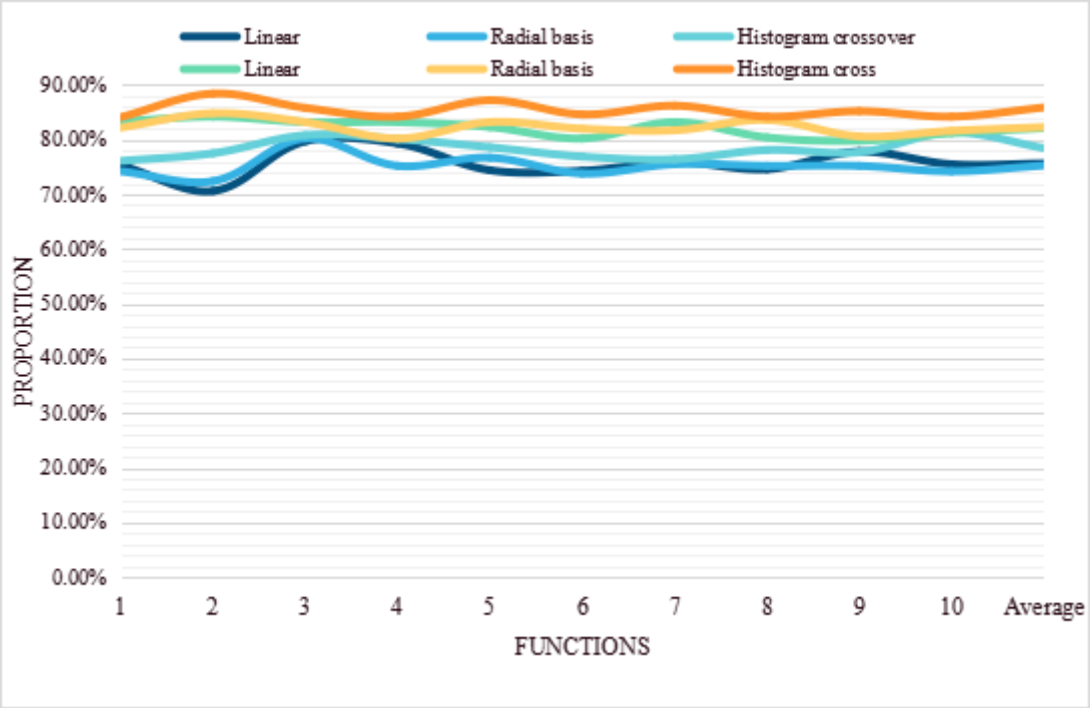

Figure 2. Analysis of the recognition rate results for cashmere wool fibres

It can be seen from the above figure 2 that after adding the spatial pyramid, the recognition rate of cashmere and wool has been significantly improved compared with the original bag model method, and the recognition rate obtained by using the histogram cross kernel function is higher than other kernel functions. As shown in Table 2 and Figure 3:

Table 2. Recognition rates under three different kernel functions

Kernel functions	Space Pyramid Improvement	Wordpack model
Linear kernel	0.82	0.76
RBF kernels	0.83	0.75
Histogram cross kernels	0.87	0.78

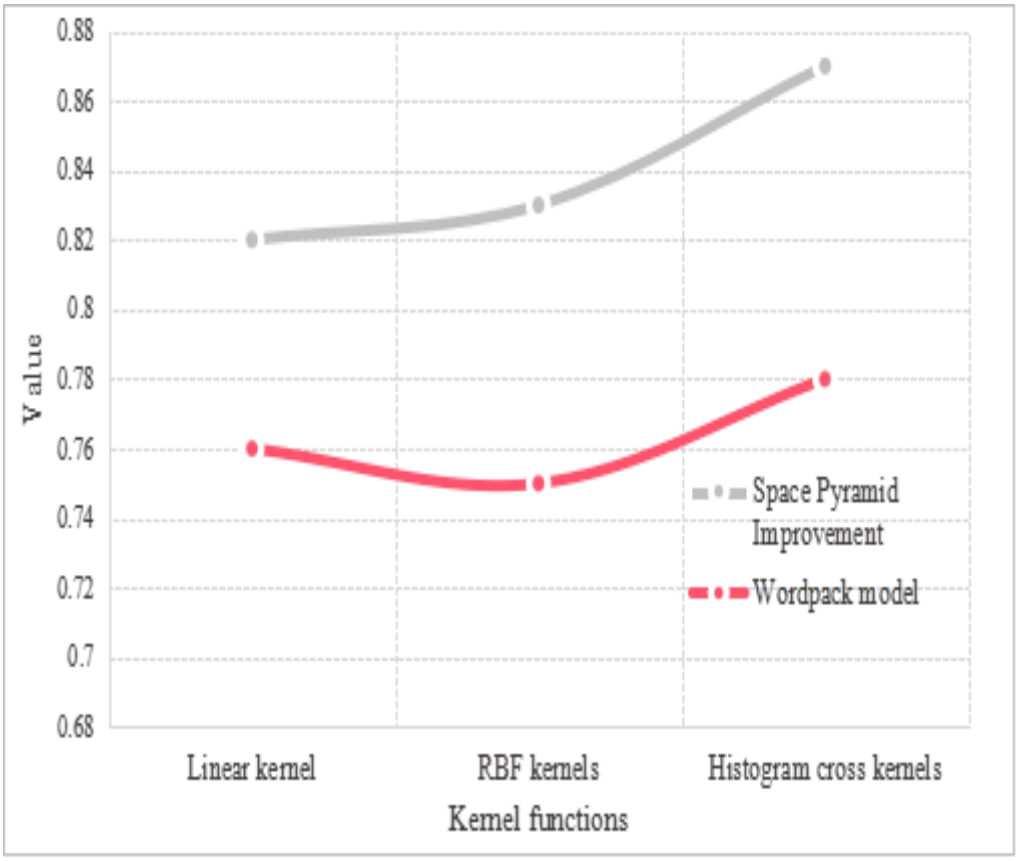

Figure 3. Comparison of recognition rates under three different kernel functions

Therefore, the histogram cross kernel function should be selected here. The formula is as follows:

Histogram cross kernel function:

$$K(x, y_i) = \sum_{k-1}^{n} \min(x_k, y_k) \tag{3}$$

Where x and yi are two arbitrary eigenvectors, xk and yk are the eigenvalues of the nth dimension of x and yi respectively, and the dimensions of the m-dimension eigenvector. It can be seen from Formula 4-8 that the calculation of histogram cross kernel function is relatively simple, which improves the operation efficiency.

2) The recognition rate obtained by different L values is shown in Table 3 below:

Table 3. Identification rates of cashmere wool fibres for different L-value parameters

L-value	Linear	RBF	Histogram crossover
1	79.3%	78.2%	80.1%
2	82.2%	81.8%	84.4%
3	83.2%	83.2%	86.8%
4	83.2%	83.4%	87.0%
5	83.3%	83.2%	87.1%

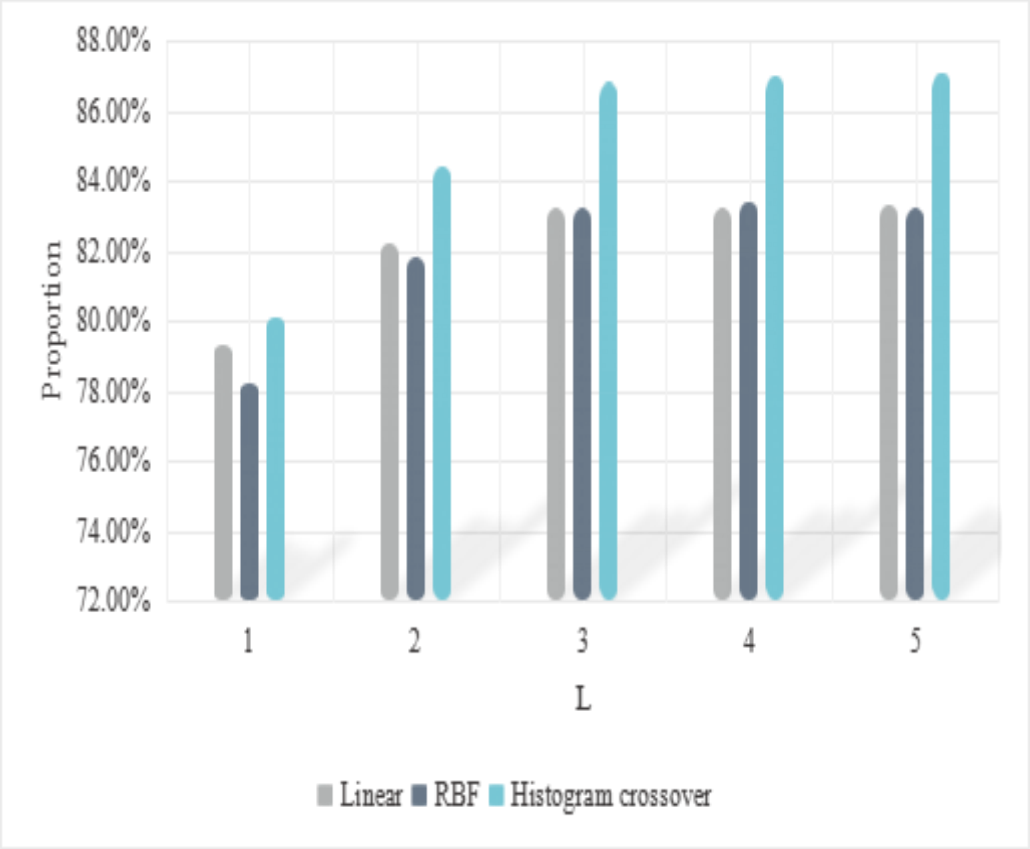

Figure 4. Comparison of the recognition rate of cashmere wool fibres with different L-value parameters

As shown in Figure 4, when the L value increases from 1 to 3, the recognition rate obtained by using these three different kernel functions will be significantly improved, but when the L value increases from 3 to 5, there is no significant difference in the recognition rate. With the increase of the L value, the operation efficiency of the computer will be significantly slower, and the hardware requirements for the computer will be higher; therefore, considering comprehensively, it is best to take L value as 3.

3) The measurement results of fiber diameter are shown in Table 4 below:

Table 4. Results of fibre diameter measurements for cashmere wool

Category	Cashmere diameter (um)	Wool diameter (um)
Mean P(um)	16.86	19.65
Maximum value(um)	25.48	30.65
Min(um)	12.16	13.56
Mean Variance	2.41	2.76
Coefficient of Variation (CV)	0.23	0.30

From the corresponding diameter values in Table 4, it can be seen that the average, maximum and minimum diameters of cashmere are smaller than those of wool. And the CV value of cashmere diameter is slightly smaller than that of wool diameter, which indicates that the difference between cashmere diameters is smaller than that of wool, and the diameter distribution range of wool is larger than that of wool.

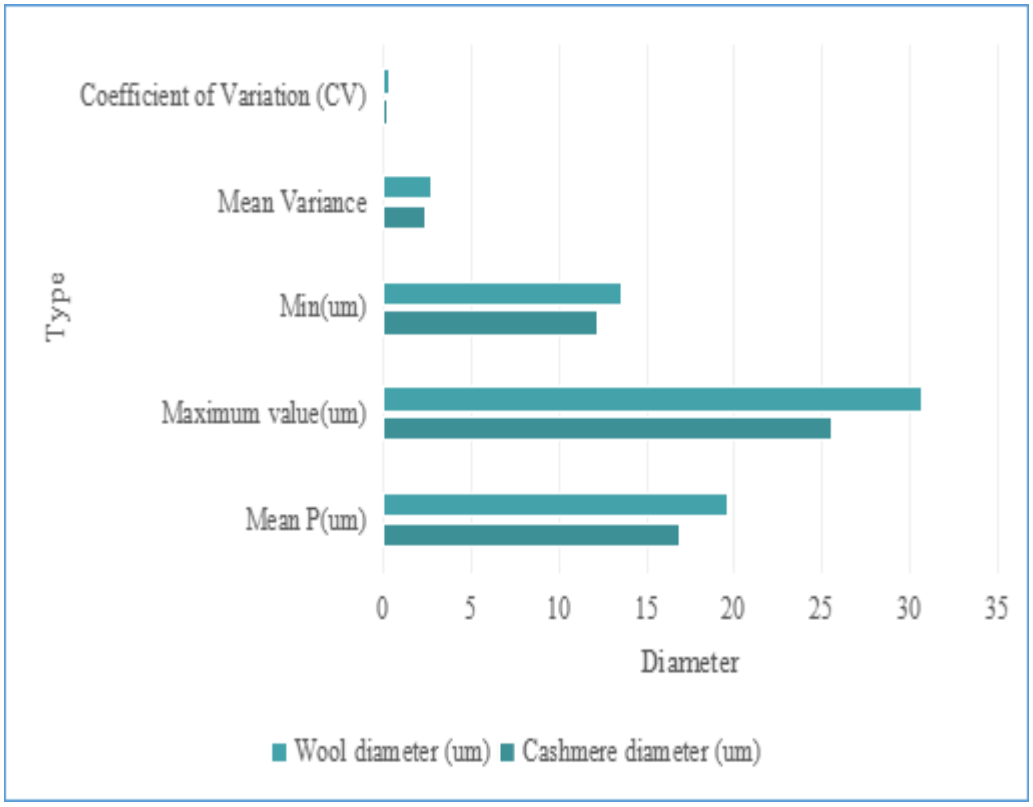

Figure 5. Comparison of fibre diameter measurements for cashmere wool

It can be seen from the above figure 5 that the diameter distribution of cashmere and wool is in normal distribution. It can be clearly seen that the diameter value of cashmere is smaller than that of wool, and the distribution is more concentrated than that of wool.

4.2. Experimental Analysis

According to the above, the K value of 500 is finally selected, the kernel function type is histogram cross kernel function, the pyramid layer L value is set to 3, and more than 1100 cashmere and wool training sets and more than 400 test sets are selected. At this time, the recognition rate is good. Through many experiments, it is found that the average recognition rate of cashmere and wool is about 86%, that is, it fluctuates around 86% each time. Through many experimental comparisons, the maximum, minimum and mean values are shown in Table 1 below:

Table 5. Recognition rate results for cashmere wool fibres

Category	Cashmere recognition rate	Wool recognition rate
Mean value	86.3%	86.1%
Maximum value	88.2%	87.6%
Min	85.3%	84.8%
Mean Variance	2.28	2.44

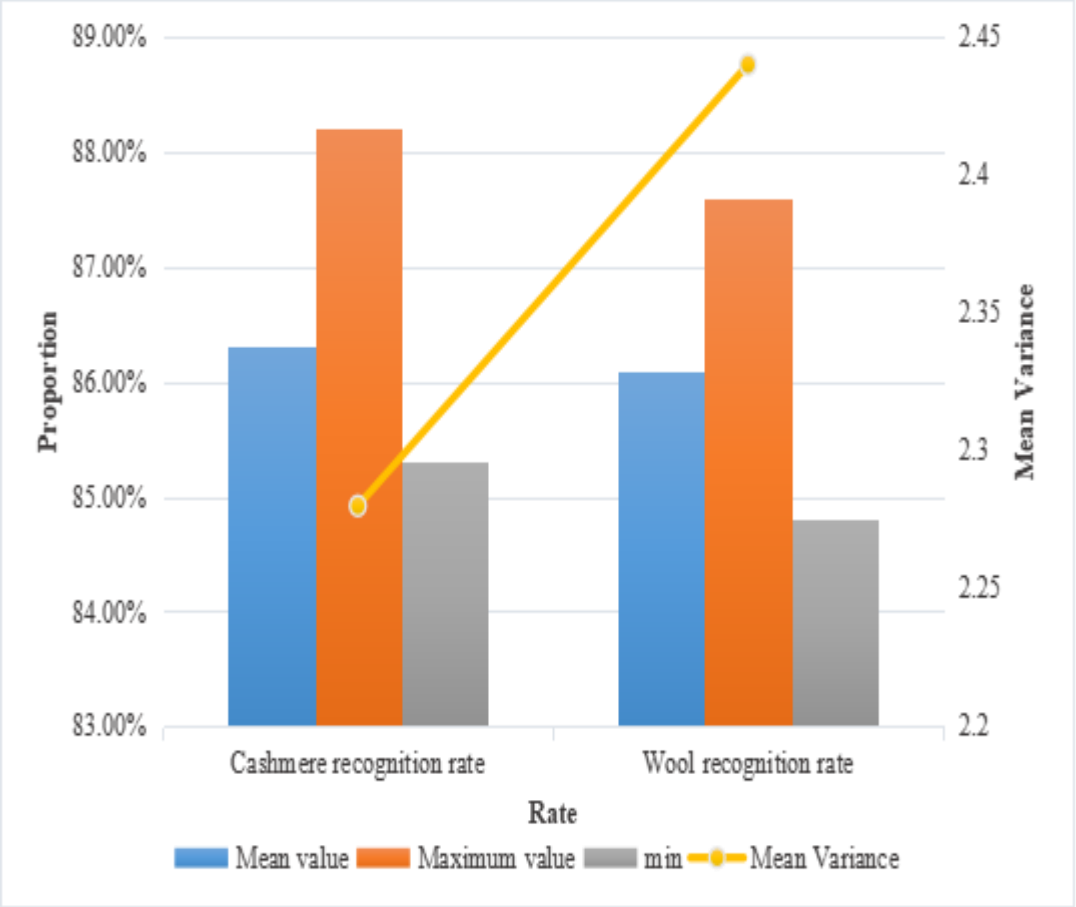

Figure 6. Comparison of recognition rate results for cashmere wool fibres

It can be seen from the above figure 6 that the maximum recognition rate of the word package model method is more than 88%. Through many experiments, the average recognition rate can reach 86%, and the results of each experiment fluctuate by 86%. The mean square deviation of the results of many experiments is 2.32, which is relatively small, indicating that the method has good robustness; In addition, this method can be used in the automatic detection system, and can be applied to the case of very large data sets. The detection speed is also relatively fast. Because the images used are from the optical microscope, its cost is relatively low, and it is easy to be popularized to practice. Therefore, this method can be used as a more effective method to identify cashmere and wool.

5. Conclusion

Animal fiber identification technology is a detection technology based on subjective judgment of human experience. Due to the variation of ecological environment, feeding conditions and animals themselves, the scale structure of animal fibers has undergone uncertain changes. This brings many difficulties to the daily inspection work. The author's unit has been engaged in the identification of animal fibers for many years, and often meets with variant cashmere fibers in the work. When encountering this problem, multiple testers are required to jointly identify, which seriously affects the work efficiency. The atlas database can realize the sharing of image information within the unit or even in many domestic units, filling the gap in this field at home and even in the world, which is of great significance to improve work efficiency.

Funding

This article is not supported by any foundation.

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Conflict of Interest

The author states that this article has no conflict of interest.

References

[1] Thai Doan Chuong, V. Jeyakumar, Guoyin Li, Daniel Woolnough: Exact SDP reformulations of adjustable robust linear programs with box uncertainties under separable quadratic decision rules via SOS representations of non-negativity. J. Glob. Optim. 81(4): 1095-1117 (2021). https://doi.org/10.1007/s10898-021-01050-x

[2] Beverly P. Woolf: Introduction to IJAIED Special Issue, FATE in AIED. Int. J. Artif. Intell. Educ. 32(3): 501-503 (2022). https://doi.org/10.1007/s40593-022-00299-x

[3] Kazim Yildiz: Identification of wool and mohair fibres with texture feature extraction and deep learning. IET Image Process. 14(2): 348-353 (2020). https://doi.org/10.1049/iet-ipr.2019.0907

[4] Stefano V. Albrecht, Michael J. Wooldridge: Multi-agent systems research in the United Kingdom. AI Commun. 35(4): 269-270 (2022). https://doi.org/10.3233/AIC-229003

[5] Anna Gautier, Michael J. Wooldridge: Understanding Mechanism Design - Part 3 of 3: Mechanism Design in the Real World: The VCG Mechanism. IEEE Intell. Syst. 37(1): 108-109 (2022). https://doi.org/10.1109/MIS.2021.3129085

[6] Liron David, Avishai Wool: Rank estimation with bounded error via exponential sampling. J. Cryptogr. Eng. 12(2): 151-168 (2022). https://doi.org/10.1007/s13389-021-00269-4

[7] Vess L. Johnson, Richard W. Woolridge, Angelina I. T. Kiser, Katia Guerra: The Impact of Coproduction Resentment on Continuation Intention. J. Comput. Inf. Syst. 62(2): 410-421 (2022). https://doi.org/10.1080/08874417.2021.1971578

[8] Lydia Barnes, Erin Goddard, Alexandra Woolgar: Neural Coding of Visual Objects Rapidly Reconfigures to Reflect Subtrial Shifts in Attentional Focus. J. Cogn. Neurosci. 34(5): 806-822 (2022). https://doi.org/10.1162/jocn_a_01832

[9] Erin Goddard, Thomas A. Carlson, Alexandra Woolgar: Spatial and Feature-selective Attention Have Distinct, Interacting Effects on Population-level Tuning. J. Cogn. Neurosci. 34(2): 290-312 (2022). https://doi.org/10.1162/jocn_a_01796

[10] Amanda K. Robinson, Anina N. Rich, Alexandra Woolgar: Linking the Brain with Behavior: The Neural Dynamics of Success and Failure in Goal-directed Behavior. J. Cogn. Neurosci. 34(4): 639-654 (2022). https://doi.org/10.1162/jocn_a_01818

[11] Daniel Woolnough, Niroshan Jeyakumar, Guoyin Li, Clement T. Loy, Vaithilingam Jeyakumar: Robust Optimization and Data Classification for Characterization of Huntington Disease Onset via Duality Methods. J. Optim. Theory Appl. 193(1): 649-675 (2022). https://doi.org/10.1007/s10957-021-01835-w

[12] Ayla Stein Kenfield, Liz Woolcott, Santi Thompson, Elizabeth Joan Kelly, Ali Shiri, Caroline Muglia, Kinza Masood, Joyce Chapman, Derrick Jefferson, Myrna E. Morales: Toward a

definition of digital object reuse. Digit. Libr. Perspect. 38(3): 378-394 (2022). https://doi.org/10.1108/DLP-06-2021-0044

[13] James Gale, Max Seiden, Deepanshu Utkarsh, Jason Frantz, Rob Woollen, Çagatay Demiralp: Sigma Workbook: A Spreadsheet for Cloud Data Warehouses. Proc. VLDB Endow. 15(12): 3670-3673 (2022). https://doi.org/10.14778/3554821.3554871

[14] Maciej Buze, Thoms E. Woolley, L. Angela Mihai: A Stochastic Framework for Atomistic Fracture. SIAM J. Appl. Math. 82(2): 526-548 (2022). https://doi.org/10.1137/21M1416436

[15] Shan Ma, Matthew J. Woolley, Ian R. Petersen: Synthesis of Linear Quantum Systems to Generate a Steady Thermal State. IEEE Trans. Autom. Control. 67(4): 2131-2137 (2022). https://doi.org/10.1109/TAC.2021.3079291

[16] Jean-Michel Fahmi, Craig A. Woolsey: Port-Hamiltonian Flight Control of a Fixed-Wing Aircraft. IEEE Trans. Control. Syst. Technol. 30(1): 408-415 (2022). https://doi.org/10.1109/TCST.2021.3059928

[17] Marcin Waniek, Tomasz P. Michalak, Michael J. Wooldridge, Talal Rahwan: How Members of Covert Networks Conceal the Identities of Their Leaders. ACM Trans. Intell. Syst. Technol. 13(1): 12:1-12:29 (2022). https://doi.org/10.1145/3490462

[18] Thomas A. Woolman, Philip Lee: Effects of Deep Learning Technologies on Employment in the Field of Digital Communication Systems. Int. J. Innov. Digit. Econ. 12(4): 35-42 (2021). https://doi.org/10.4018/IJIDE.2021100103