Script Identification of Central Asian Printed Document Images based on Nonsubsampled Contourlet Transform

Xing-kun Han, Alimjan Aysa, Hornisa Mamt, and Kurban Ubul*

***Abstract*—Document images of various scripts must be identified and processed in today's international environment. As the front-end technology of Optical Character Recognition (OCR), script identification is an indispensable part of automatic document image analysis. Aiming at the nature of rich texture features of document images, a 3-level Nonsubsampled Contourlet Transform (NSCT) was used to extract 30- dimensional texture features in this paper. A Support Vector Machine (SVM) and K Nearest Neighbor (KNN) classifier were used for classification. A total of 10,000 document images in 10 kinds of Central Asian scripts—Arabic, Russian, Tibetan, Chinese, Uyghur, English, Mongolian, Kyrgyzstan, Kazakhstan, and Turkish—were classified. The identification efficiency of SVM and KNN was analyzed and compared, with the result that the SVM classifier obtained 99.5% average accuracy, a higher accuracy than KNN, during the experiment. The validity of the proposed method was proved by comparing the Wavelet Transforms (WT) and Local Binary Patterns (LBP) of these two script-identification methods.**

***Index Terms*—Identification of Central Asian scripts, texture feature, nonsubsampled contourlet transform, support vector machine**

I. Introduction

With the gradual development of science and technology and the concept of environmental protection, the paperless office has been promoted around the world. Scanning the paper version of a document into images can greatly reduce storage-space requirements, and is more conducive to the preservation and processing of document information. OCR is a process of scanning paper documents, recognizing characters, and creating a corresponding electronic document. Its development extends the application of document images to the fields of classification, retrieval, and translation etc. Most OCR systems need to know the type of script a document uses before recognizing a conversion character. This tedious, time-consuming task must be performed manually when processing large volumes of document images. Therefore, it is important to research the automatic identification of the script of a document image, which can greatly reduce the processing time of an OCR system and make it more intelligent.

There are currently two main script-identification methods, one based on structural features and the other on texture features [1], [2]. The method based on structural features [3], [4], [5], [6] has high accuracy. However, since such methods are based on text lines or characters and word levels, the document image must be precisely segmented. This leads to poor robustness with respect to noise, tilt, and so on. Image texture is one kind of perception of the human visual system to the surface phenomena of objects. It is one of the important characteristics people use to describe and distinguish different objects. Characters display various combinations of writing characteristics, stroke directions, and sparseness. These features will appear in a document image with different texture features. Therefore, texture features can be used for script identification of a document image. The method of script identification based on texture features [7], [8], [9], [10], [11], [12], [13], [14], [15] extracts the features of the entire document image block. This method has good robustness to noise and tilt, and has become a research topic.

G. S. Peake and T. N. Tan [7] first proposed the use of texture features for script identification in 1997, using the Gabor filter method for English, Korean, Greek, and 7 other kinds of script in the case of a small sample set. The accuracy rate reached 95%. L. Zeng et al. [8] proposed a multi-scale wavelet transform to extract the multi-scale wavelet energy of a document image as a texture feature. They experimented with 6 kinds of scripts, including Chinese, English, and Japanese, with average accuracy over 90%. A. Busch et al. [9] experimented on 8 kinds of scripts, such as English, Hebrew, and Hindi, using the texture feature based on the Gray-level Co-occurrence Matrix (GLCM), with an identification error rate of about 9%. L. Gu et al. [10] proposed the application of a steerable pyramid transformation to script identification, and the steerable pyramid energy statistic texture feature of the document image was extracted. Chinese, Russian, Burmese, and 7 other kinds of script were identified, and the

Manuscript received March 27, 2017; revised June 18, 2017. This work was supported by the National Natural Science Foundation of China (No. 61363064, 61563052, 61163028), College Scientific Research Plan Project of Xinjiang Uyghur Autonomous Region (No. XJEDU2013I11), and Special Training Plan Project of Xinjiang Uyghur Autonomous Region’s Minority Science and Technological Talents (No. 201323121).

Xing-kun Han. Author is with the School of Information Science and Engineering, Xinjiang University, No.666 Shengli Road, Urumqi, Xinjiang, 830046, China. (e-mail: xkhan0810@163.com)

Alimjan Aysa. Author is with the Network and Information Technology Center, Xinjiang University, No.666 Shengli Road, Urumqi, Xinjiang, 830046, China. (e-mail: alim@xju.edu.cn).

Hornisa Mamt. Author is with the School of Information Science and Engineering, Xinjiang University, No.666 Shengli Road, Urumqi, Xinjiang, 830046, China. (e-mail: hornisa1016@163.com).

Kurban Ubul. Corresponding author is with the School of Information Science and Engineering, Xinjiang University, No.666 Shengli Road, Urumqi, Xinjiang, 830046, China. (Corresponding author phone: +86-991-858-2558; fax: +86-991-858-0288; e-mail: kurbanu@xju.edu.cn).

average identification rate reached 95.3%. M. A. Ferrer et al. [11] applied LBP to 10 kinds of script, including Bengali, Roman, and Urdu, with an identification rate of over 99%.

Although script-identification technology based on texture features has been developed for many years and achieved good results, many methods are aimed at scripts of certain regions and countries, and cannot be applied to all the world's scripts. Especially for the Central Asian region, there is no such research. With the development of the integration of the world and the Belt and Road advance, economic and cultural exchanges between China and other Central Asian countries will become more frequent. It will generate many paper documents to be stored and processed. As the front-end technology of OCR, the script identification of printed documents in Central Asia has become an urgent research topic.

Based on the characteristics of the similarities among the scripts in Central Asia, this paper proposes a new method for script identification based on NSCT [16] and SVM. The texture features of document images are extracted by NSCT and classified by SVM. To verify the effectiveness of this method, WT [9] and LBP [11] are used to compare with it. The effectiveness of the method is also verified by comparing other experimental results of previous work.

II. Nonsubsampled Contourlet Transform

NSCT is developed by a contourlet transform [17]. It has the same multi-directional, multi-scale, and anisotropic features as a contourlet transform, and it also borrows the à trous algorithm [18]. NSCT does not directly perform down-sampling operation on the decomposition and reconstruction process, but perform up-sampling operation on the decomposition filter and synthesis filter. This gives NSCT the characteristics of shift-invariance, and the phenomena of spectrum-aliasing and leakage do not occur when processing the image. Moreover, the detail sub-image and the original image are the same size, so the original image information can be preserved well. It can retain the original image information and more accurately describe the details of image. Therefore, compared with other texture-feature extraction methods, NSCT is more suitable for the script identification of document images with rich texture features.

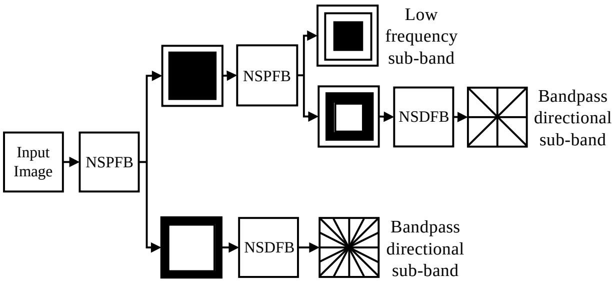


Fig. 1. The nonsubsampled contourlet transform

NSCT is composed of two parts: a nonsubsampled pyramid filter bank (NSPFB) and a nonsubsampled direction filter bank (NSDFB). Its structure is shown in Fig. 1. Both NSPFB and NSDFB are composed of dual-channel nonsubsampled filter banks. The former can transform the input signal into multi-scale and multi-resolution signals, while the latter can decompose the band-pass signal into 2^n (n = 0, 1, 2...) directional sub-bands and give the transform multi- directionality. The nonsubsampled dual-channel filter banks must satisfy the Bezout identity:

$$W_0(z)Y_0(z)+W_1(z)Y_1(z)=1 \tag{1}$$

where $W_0(z)$ and $Y_0(z)$ are the frequency responses of the low- and high-pass decomposition filters, respectively, and $W_1(z)$ and $Y_1(z)$ are the respective frequency responses of the low- and high-pass synthesis filters.

In the process of NSCT, the input image is first decomposed by NSPFB to generate high- and low-frequency sub-bands. The high-frequency sub-band is then decomposed by NSDFB to obtain sub-bands of different directions, and then the low-frequency part continues through the NSPFB to perform the next nonsubsampled tower-decomposition. In this way, the multi-scale and multi-directional decomposition sub-images of the input image can be obtained by repeating the above operations for the low-frequency sub-bands of each layer.

III. Script Identification based on NSCT and SVM

Script identification of document images of Central Asia in this paper mainly include the following four parts: creation of document image database, preprocessing, feature extraction, and classification. The specific process is shown in Fig. 2.

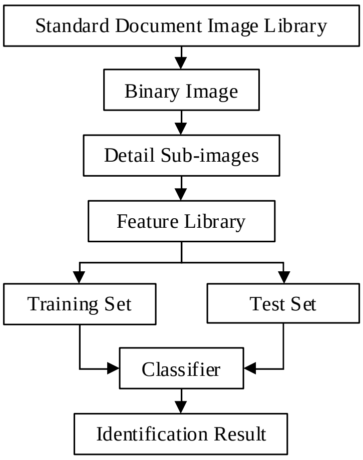


Fig. 2. Flowchart of Central Asian scripts identification

First, the document images are formed by scanning paper documents such as books and magazines etc. A standard document image based on the same size is created by cutting the full-page document image. Then, the binary image is formed by preprocessing, such as graying, binarization, and denoising. Then, the binary image is decomposed by NSCT to produce a series of sub-images. The features are extracted from each sub-image to build the feature library. Finally, the feature library is divided into a training set and test set, which are trained and classified by a classifier.

A. Preprocessing

In the preprocessing stage, the image is grayed by the weighted-average method and binarized using the global adaptive thresholding method. The median filter is then used to remove the salt-and-pepper noise generated during the scanning process. Their purpose is to improve the classification accuracy by eliminating the impact of the color of the document image and the noise.

B. Feature Extraction

The NSCT with *l*-scale decomposition of document image *I* can be expressed as

$$I = a_j + \sum_{j=1}^{l} \sum_{k=1}^{n} W_j^k \tag{2}$$

where a_j denotes the low-frequency sub-band at the *j-th* scale and W_j^k refers to the sub-band of the *k-th* direction at the *j-th* scale.

To reduce the influence of the font format, size, and stroke break on accurate classification in the document image, the preprocessed image was decomposed in 3-scale, and 2^1, 2^2, and 2^3 directions from low- to high-scale were respectively decomposed. In this way, a low-frequency sub-image and 14 sub-images with the same size in different directions at different scales were obtained. As shown in Fig. 3, the transformed sub-image is saved as a cell array.

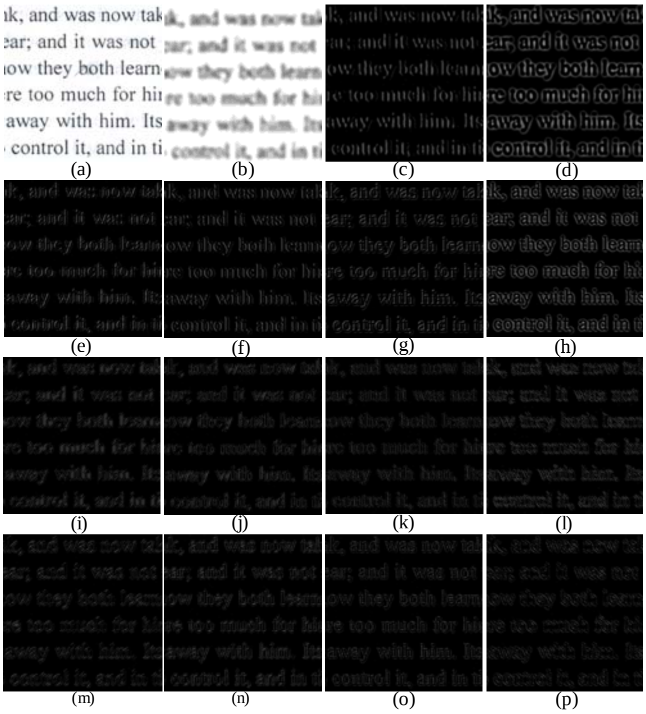


Fig. 3. Sub-images after nonsubsampled contourlet transform (a) original image (b) low-frequency sub-image (c)-(d) first scale sub-images (e)-(h) second scale sub-images (i)-(p) third scale sub-images

The mean and variance of each sub-image were taken to be the texture features needed for the experiment:

$$\mu_{\mathrm{j}}^{k} = \frac{1}{MN} \sum_{x=1}^{M} \sum_{y=1}^{N} W_j^k(x,y) \tag{3}$$

$$\sigma_j^{k^2} = \frac{1}{MN-1} \sum_{x=1}^{M} \sum_{y=1}^{\mathrm{N}} \left(W_j^k(x,y) - \mu_j^k \right)^2 \tag{4}$$

where *M* and *N* are the size of image, and their units are pixels, *j* is the decomposition scale, and *k* is the decomposition direction. From the above, we obtain a 30-dimensional feature vector:

$$X = \begin{bmatrix} \mu, \sigma^2, \mu_1^1, \sigma_1^{1^2}, \mu_1^2, \sigma_1^{2^2}, \mu_2^1, \sigma_2^{1^2}, \mu_2^2, \sigma_2^{2^2}, \mu_2^3, \\ \sigma_2^{3^2}, \mu_2^4, \sigma_2^{4^2}, \mu_3^1, \sigma_3^{1^2}, \mu_3^2, \sigma_3^{2^2}, \mu_3^3, \sigma_3^{3^2}, \mu_3^4, \\ \sigma_3^{4^2}, \mu_3^5, \sigma_3^{5^2}, \mu_3^6, \sigma_3^{6^2}, \mu_3^7, \sigma_3^{7^2}, \mu_3^8, \sigma_3^{8^2} \end{bmatrix} \tag{5}$$

C. Training and Classification

SVM, as proposed by Vapnik, can be used for pattern recognition and nonlinear regression [19]. Its main idea is to map the sample space into a high-dimensional space through a nonlinear mapping and establish a classifying hyperplane as the decision surface. It transforms the problem of indivisibility in the original low-dimensional sample space to a linear space in the high-dimensional feature space. The computational complexity is reduced by the introduction of a kernel function. Because SVM has many advantages in solving nonlinear and small sample-classification problems, it has been widely used in text classification, pattern recognition, and other fields. In this paper, SVM toolbox Libsvm, developed by Lin Chin-Jenkai, is used to classify document images. Cross-validation is used to obtain the best parameters, and the radial basis function shown in equation (6) is chosen as the kernel function of SVM.

$$K(x_i, x_j) = \exp\left(\frac{-\left\| x_i - x_j \right\|^2}{2\sigma^2} \right) \tag{6}$$

To facilitate the data processing and accelerate the convergence speed, we need to normalize the training-set eigenvectors and test-set eigenvectors. In this paper, normalization is performed in the [-1, 1] interval, which maps as follows:

$$f: x \rightarrow y = 2 \times \frac{x - x_{min}}{x_{max} - x_{min}} + (-1) \tag{7}$$

where $x, y \in R^n$, $x_{min}=min(x)$, $x_{max}=max(x)$**.**

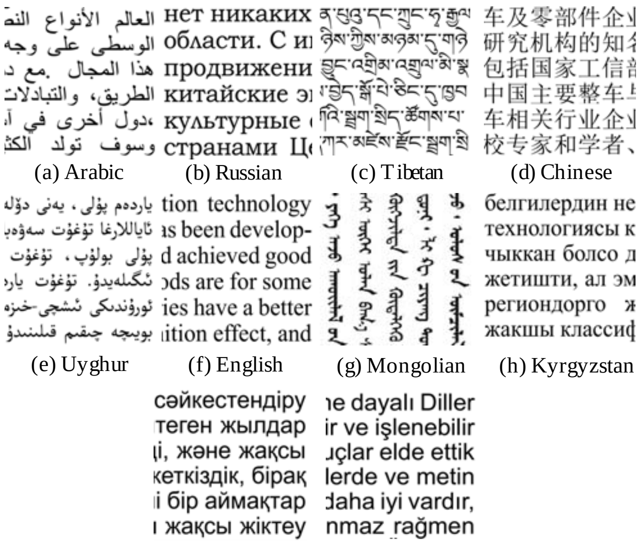
To verify the superiority of SVM as the classification algorithm, the KNN algorithm, which performs well in text classification, is selected for a comparative experiment. Although KNN has many improved algorithms [20], but the core idea of them is same that if most of the K nearest neighbors of a sample in a feature space belong to a certain class, the sample also belongs to the class and has the characteristics of the samples in that class. The method classifies a sample based on the category of the nearest one or several samples. To use the KNN classification algorithm, it is critical to choose suitable value of K and the distance. If K is too small, there are too few neighbors, which will amplify the interference of noise data and reduce the classification accuracy. And if the K is chosen too large, the data that is not similar to the sample to be tested will be included, which will result in increased noise and a reduced identification rate. In this paper, the best value of K is set to 3 and the distance of "cityblock" is chosen by cross-validation in the experiment, which is equation (8):

$$d_{12} = \sum_{k=1}^{n} \left| x_{1k} - x_{2k} \right| \tag{8}$$

IV. Experiment Results and Analysis

A. Experiment Database

For this paper, the books and magazines of 10 commonly used scripts in China and Central Asia were scanned to collect document images to build a database. The database includes Chinese, English, Uyghur, Mongolian, Tibetan, Russian, Arabic, Kyrgyzstan, Kazakhstan, and Turkish. In the experimental stage, a full-page image is cut into the same size 256*256 dpi image. The experimental database containing 1,000 images in each script, and a total of 10,000 document images are included. Fig. 4 shows some samples in the experimental database.


нет никаких
области. С и
продвижени
китайские э
культурные
странами Ц
车及零部件企
研究机构的知
包括国家工信
中国主要整车
车相关行业企
校专家和学者、
tion technology
s been develop-
d achieved good
ods are for some
ies have a better
ition effect, and
белгилердин не
технологиясы к
чыккан болсо д
жетишти, ал эм
региондорго ж
жакшы классиф
сәйкестендіру
теген жылдар
і, және жақсы
кеткіздік, бірақ
і бір аймақтар
жақсы жіктеу
e dayalı Diller
ir ve işlenebilir
çlar elde ettik
lerde ve metin
daha iyi vardır,
nmaz rağmen

(a) Arabic (b) Russian (c) Tibetan (d) Chinese

(e) Uyghur (f) English (g) Mongolian (h) Kyrgyzstan

(i) Kazakhstan (j) Turkish

Fig. 4. Some samples in the experimental database

B. Filter Bank Selection

There are many types of filters that make up NSPFB and NSDFB. Different filter combinations will form different NSCTs. Whether the filter bank can extract the texture features of the document image accurately and effectively will affect the final identification accuracy. And the time required for feature extraction will directly affect the identification efficiency. Therefore, to improve the performance of the entire script identification system, selecting the appropriate combination of filters is important. The filter bank with higher identification accuracy and shorter feature-extraction time than others will be the best choice.

To select the optimal combination, we tested different kinds of filter banks to compare their classification accuracy rate (R) and the time of feature extraction (T). The classifier used was SVM. The experiment of each combination was conducted 10 times. The average of the 10 experimental results was taken as the final result. Half of the document images were randomly selected from the document image database as the training set, and the remaining were taken as the test set. Test results are shown in Table I, in which “9-7”, “maxflat”, “pyr”, and “pyrexc” are pyramid filter, and “haar (ha)”, “vk”, “ko”, “kos (ks)”, “lax (la)”, “sk”, “cd”, “pkva (pk)”, “dvmlp (dp)”, “sinc (si)”, “dmaxflat4 (d4)”, “dmaxflat5 (d5)”, “dmaxflat6 (d6)”, and “dmaxflat7 (d7)” are direction filters. The accuracy rate is defined in equation (9):

$$R = \frac{m_1}{m_2} \tag{9}$$

where m_1 is the total number of correctly identified samples, and m_2 is the number of test samples.

It can be seen from Table Ⅰ that the combination with “dvmlp” filter as the direction filter has the highest identification accuracy, but the feature-extraction phase takes a longer time. The identification accuracy of the combination with “haar” filter as the direction filter is 0.7% lower than the average of the combination with the “dvmlp” filter as the direction filter. However, its feature-extraction phase takes the shortest time. Its time of feature extraction is 1/6 that of the combination with “dvmlp” filter as the direction filter. Although there are other combinations with higher identification accuracy, such as the accuracy rate of “9-7 + dmaxflat7” combination reached 99.58%. But its feature extraction time is 6.29s, which will greatly affect the identification efficiency of the whole system. Therefore, considering the efficiency of the system, this paper uses “9-7 + haar” filter bank.

TABLE I
EXPERIMENTAL RESULTS OF DIFFERENT NSCT FILTER BANKS

	9-7		maxflat		pyr		pyrexc	
	R(%)	T/s	R(%)	T/s	R(%)	T/s	R(%)	T/s
ha	99.58	0.08	99.57	0.15	99.53	0.08	99.52	0.08
vk	99.46	0.14	99.33	0.20	99.28	0.12	99.28	0.13
ko	99.40	0.13	99.30	0.19	99.26	0.11	99.26	0.12
ks	99.53	0.13	99.36	0.19	99.37	0.11	99.37	0.12
la	99.58	1.04	99.50	1.15	99.38	1.06	99.40	1.05
sk	99.55	0.33	99.41	0.37	99.36	0.31	99.31	0.32
cd	99.49	0.29	99.29	0.33	99.22	0.27	99.25	0.27
pk	99.62	4.00	99.49	4.23	99.52	4.20	99.52	4.09
dp	99.64	0.51	99.65	0.56	99.61	0.50	99.59	0.49
si	99.60	3.13	99.57	3.18	99.58	3.08	99.57	3.01
d4	99.57	2.01	99.45	2.09	99.46	2.12	99.45	2.10
d5	99.60	3.07	99.47	3.17	99.44	3.11	99.46	3.09
d6	99.59	4.40	99.48	4.46	99.46	4.53	99.50	4.53
d7	99.58	6.29	99.47	6.40	99.47	6.14	99.48	6.21

C. Comparative Experiments of Different Methods

To reflect the classification result of the SVM algorithm for script identification, this paper uses the SVM and KNN classification algorithms, respectively. In the experiments, 10% to 90% of the document images of each script were randomly selected from the image database as a training set. The remaining images were used as a test set for different experiments. Each experiment was repeated 10 times. The

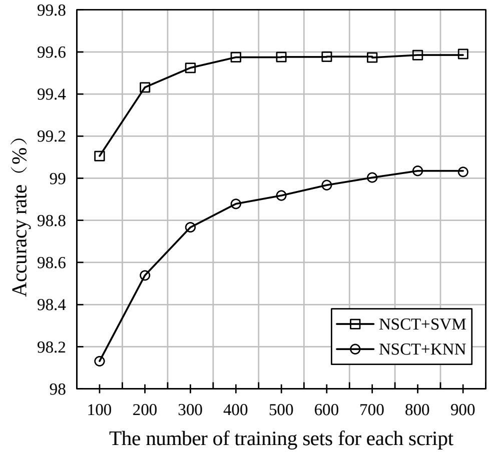


Fig. 5. Experiment results of different classifiers

average of the experimental results was taken as the final result. The experimental results are shown in Fig. 5.

Experimental results show that the identification rate of the two classifiers increased steadily with the increase in the number of the training samples. The identification accuracy based on SVM was higher than for KNN overall. For the SVM classifier, the identification rate was 99.11% when the number of training sample was 1,000, i.e., the sample was 10% of each script image database. However, when the number of training sample was 8,000, the identification rate based on the KNN classifier reached its highest value, 99.03%. Overall, the effect of classification based on SVM was significantly better than that based on KNN.

To verify the validity of the proposed method, WT [9] and LBP [11] were compared. In the comparative experiment, different number of document images were randomly selected from the document image database as a training set, and the remaining were used as a test set. In the same way, each experiment was repeated 10 times. The average of the experimental results was taken as the final result. The experimental results are shown in Fig.6.

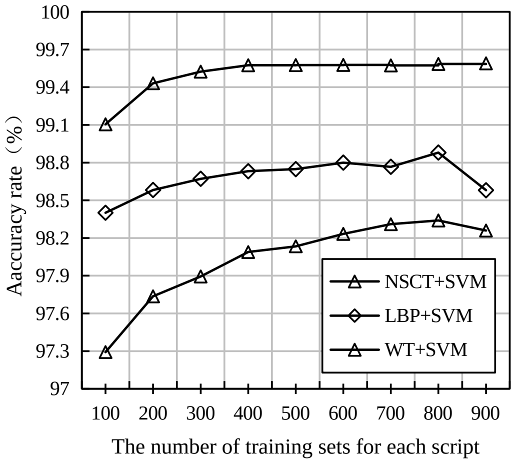


Fig. 6. Experimental results of different methods

It can be seen from Fig. 6 that the accuracy of the proposed method in this paper was a significant improvement over the other two methods. The identification accuracy of the proposed method was more than 99% when the training sample was 1/10 of all samples. This indicates that the method has good generalization ability. Compared with WT and LBP, the average accuracy increased by 1.5% and 0.82%, respectively, but the feature-extraction time (Table Ⅱ) only increased by 0.049s, and the identification efficiency was higher.

TABLE Ⅱ
FEATURE-EXTRACTION TIME OF DIFFERENT METHODS

Method	Feature extraction time (s)
WT	0.038
LBP	0.038
NSCT	0.087

The precision rate *TP* and the recall rate *TR* are widely used in the field of information retrieval and statistical classification to evaluate the quality of a system. We use these two indicators to evaluate the identification effect of the proposed method again. They are defined as:

$$TR = \frac{n_r}{n_t} \tag{10}$$

$$TP = \frac{n_r}{n_r + n_w} \tag{11}$$

Where, for a given category to be tested, n_r is the number of samples correctly identified in this category, n_w is the number of samples that are misidentified as this category, and n_t is the total number of samples used for testing in this category.

In the experiment, half of the document images were randomly selected from Chinese (Ch), English (En), Uyghur (Uy), Mongolian (Mo), Tibetan (Ti), Russian (Ru), Arabic (Ar), Kyrgyzstan (Ky), Kazakhstan (Ka), and Turkish (Tu) in the document image library as a training set, and the remaining were used as a test set. To minimize the effect of the differences between the samples on the classification result, a total of 10 experiments were conducted, and their average was used as the final result. The results are shown in Table Ⅲ, where AV represents the overall average value.

TABLE Ⅲ
IDENTIFICATION EFFECT OF DIFFERENT METHODS

	TR (%)			TP (%)		
	WT	LBP	NSCT	WT	LBP	NSCT
Ar	99.62	99.84	99.86	99.34	99.96	100
Ru	99.92	99.98	100	99.80	99.94	100
Ti	100	100	100	100	100	100
Ch	99.88	100	100	99.98	99.82	100
Uy	99.36	99.96	100	99.50	99.84	99.87
En	100	99.98	100	100	99.98	100
Mo	100	100	100	99.88	100	100
Ky	91.08	95.32	97.96	92.14	93.40	97.96
Tu	99.68	99.3	99.92	99.66	99.93	100
Ka	91.99	93.26	98.02	91.00	94.74	97.91
AV	98.15	98.76	99.58	98.13	98.76	99.57

From the analysis of Table Ⅲ, for texture features, which are quite different between scripts such as Arabic, Tibetan, Chinese, English, and Mongolian, each algorithm can achieve almost error-free identification. But no matter which algorithm, Kyrgyzstan and Kazakhstan will show a higher error rate than others. This is mainly because the two scripts are composed of the Cyrillic alphabet and their structures are similar, making their textures more similar. This results in lower texture feature discrimination, leading to difficulty distinguishing between the two scripts. Hence, both *TP* and *TR* are lower for them than for other scripts. However, the identification accuracy of the method used in this paper is a great improvement over WT and LBP. For example, for Kyrgyzstan, *TR* and *TP* of NSCT are both 97.96%, which are respectively 6.88% and 5.82% higher than WT, and 2.64% and 4.56% higher than LBP. For Kazakhstan, *TR* and *TP* are respectively 98.02% and 97.91% for NSCT, which are 6.03% and 6.91% higher than for WT, and 4.76% and 3.17% higher than for LBP. These results also indicate that NSCT is superior to the texture-feature extraction. Therefore, because NSCT does not perform subsampling in the transformation, so the local detail information of the sub-image is preserved completely, and the similar scripts can also obtain a good identification effect.

D. Comparison with Previous Methods

To illustrate the effectiveness of the proposed method, this paper summarizes the previous identification results of the commonly used methods for script identification. The results are shown in Table Ⅳ.

TABLE Ⅳ
THE COMPARISON RESULTS BETWEEN THE METHOD AND PREVIOUS METHODS

Author	Method	Classifier	Training set	Test set	R (%)
S. J. Lu et al. [21]	Traversing times	KNN	80	80	95.63
U. Pal et al. [3]	Water reservoir principle, profile, etc.	Binary tree	1,250	2,750	97.52
B. Mijit et al. [22]	HSV	BP	1,100	1,100	88.14
T. N. Tan [14]	Gabor	——	90	60	96.70
A. Busch et al. [9]	Wavelet energy	LDA	800	800	95.40
A. Busch et al. [9]	GLCM	LDA	800	800	90.10
M. A. Ferrer et al. [11]	LBP	SVM	250	500	95.41
P. S. Hiremath et al. [15]	Wavelet-based co-occurrence histogram	KNN	3,200	3,200	98.00
This paper	NSCT	SVM	5,000	5,000	99.58

As shown in the table, S. J. Lu et al. [21] and U. Pal et al. [3] obtained 95.63% and 97.52% accuracy rates, respectively, with different structural features and different classifiers. B. Mijit et al. [22] extracted hue, saturation, value features (HSV) of a document image and classified it with the BP network. An average 88% accuracy was obtained with 1,100 test samples. T. N. Tan [14] used a multi-channel Gabor filter to perform multiscale transformations of document images, and extracted the texture features of subgraphs. Its identification accuracy reached 96.7%. In another study [9], the wavelet energy and the GLCM of the document image were extracted as the texture features, with accuracy of 95.40% and 90.1%, respectively, obtained by a Linear Discriminant Analysis (LDA) classifier. M. A. Ferrer et al. [11] used LBP to extract the texture features of a document image and used the SVM classifier to train and test, with a 95.41% accuracy rate. Another study [15] used a wavelet transform to extract a co-occurrence histogram of document images as a texture feature. Its identification accuracy reached 98%. The method of script identification based on NSCT and SVM in this paper obtained better results than the above methods, with a 99.58% accuracy rate over 5,000 training samples and 5,000 test samples. Through the above comparative analysis, the effectiveness of the method is proved.

V. Conclusion

In this paper, a standard document image database of 10 kinds of commonly used Chinese scripts and some Central Asian scripts, including a total of 10000 document images, was built. And according to the rich nature of document image texture, the NSCT was applied to extract the texture feature of the document image to identify the script. Considering the feature-extraction time and identification accuracy, a suitable combination of pyramid and direction filters was selected for experiments. The SVM classifier was used for classification, with a better effect than the KNN classifier. It obtained an average accuracy greater than 99%. Compared with other traditional methods, such as WT and LBP, the method used in this paper had better identification accuracy for similar scripts. It is proved that this method can effectively extract the texture features of document images, and it can be applied to other scripts. Kyrgyzstan and Kazakhstan, which are similar scripts, had the highest rate of error identification in our experiments. So, the focus of future research is to combine other methods to improve the identification rate of similar scripts.

References

[1] D. Ghosh, T. Dube, and A. P. Shivaprasad, “Script Identification—A Review,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no.12, pp. 2142-2161, 2010.

[2] K. Ubul, G. Tursun, A. Aysa, D. Impedovo, G. Pirlo and T. Yibulayin, “Script Identification of Multi-Script Documents: a Survey,” *IEEE Access*, vol. 5, pp. 6546-6559, 2017.

[3] U. Pal, S. Sinha, and B. B. Chaudhuri, “Multi-Script Line identification from Indian Documents,” in *Document Analysis and Recognition, 2003. ICDAR’03. 3th International Conference on*, vol. 3163, pp. 880-884, 2003.

[4] S. M. Obaidullah, A. Mondal, and K. Roy, “Structural feature based approach for script identification from printed Indian document,” in *Signal Processing and Integrated Network, 2014 International Conference on*, pp. 120-124, 2014.

[5] M. M. Goswami and S. K. Mitra, “Classification of Printed Gujarati Characters Using Low-Level Stroke Features,” *ACM Transactions on Asian and Low-Resource Script Information Processing*, vol. 15, no. 4, pp. 1-26, 2016.

[6] R. Bashir and S. M. K. Quadri, “Entropy based Script Identification of a multilingual Document Image,” in *Computing for Sustainable Global Development, 2014 International Conference on*, vol. 7, no. 2, pp. 19-23, 2014.

[7] G. S. Peake and T. N. Tan, “Script and language identification from document images,” *Document Image Analysis*, vol. 1, pp. 10:17, 1997.

[8] L. Zeng, Y. Tang, and T. Chen, “Automatic script identification based on multi-scale wavelet texture analysis,” *Chinese Journal of Computers*, vol. 23, no. 7, pp. 699-704, 2000.

[9] A. Busch, W. W. Boles, and S. Sridharan, “Texture for script identification,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 27, no. 11, pp. 1720-1732, 2005.

[10] L. Gu, X. Ping, and J. Cheng, “A script identification method with rotational robustness,” *Chinese Journal of Image and Graphics*, vol. 15, no. 6, pp. 879-886, 2010.

[11] M. A. Ferrer, A. Morales, and U. Pal, “LBP Based Line-Wise Script Identification,” in *Document Analysis and Recognition, 2013. ICDAR’12. 12th International Conference on*, pp. 369-373, 2013.

[12] S. Chaudhari and R. M. Gulati, “Script Identification Using Gabor Feature and SVM Classifier,” *Procedia Computer Science*, vol. 79, pp. 85-92, 2016.

[13] F. K. Jaiem, S. Kanoun, and V. Eglin, “Arabic Font Identification Based on a Texture Analysis,” in *Frontiers in Handwriting Identification, 2014 International Conference on*, pp. 673-677, 2014.

[14] T. N. Tan, “Rotation invariant texture features and their use in automatic script identification,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 20, no. 7, pp. 751-756, 1998.

[15] P. S. Hiremath and S. Shivashankar, “Wavelet based co-occurrence histogram features for texture classification,” *Pattern Recognition Letters*, vol.29, pp. 1182–1189, 2008.

[16] A. Cunha, J. Zhou, and M. N. Do, “The nonsubsampled contourlet transform: theory, design, and applications,” *IEEE Transactions on Image Processing*, vol. 15, no. 10, pp. 3089-3101, 2006.

[17] M. N. Do and M. Vetterli, “The contourlet transform: an efficient directional multiresolution image representation,” *IEEE Transactions on Image Processing*, vol. 14, no. 12, pp. 2091-2106, 2005.

[18] M. J. Shensa, “The discrete wavelet transform: wedding the à trous and Mallat algorithms,” *IEEE Transactions on Signal Processing*, vol. 40, no. 10, pp. 2464-2482, 1992.

[19] M. F. Hashmi, A. R. Hambarde, and A. G. Keskar, “Robust Image Authentication Based on HMM and SVM Classifiers,” *Engineering Letters*, vol. 22, no. 4, pp.183-193, 2014.

[20] H. S. Nagendraswamy and D. S. Guru, “K-Mutual Nearest Neighbour Approach for Clustering Two-Dimensional Shapes Described by Fuzzy-Symbolic Features,” *Engineering Letters*, vol. 14, no. 1, pp. 143-153, 2007.

[21] S. J. Lu and C. L. Tan, “Automatic Detection of Document Script and Orientation,” in *Document Analysis and Recognition, 2007. ICDAR’07, 7th International Conference on*, vol. 1, pp. 237-241, 2007.

[22] B. Mijit, A. Aysa, N. Yadikar, X. K. Han, and K. Ubul, “Script Identification Based on HSV Features,” in *Pattern Recognition, 2016. CCPR’07, 7th Chinese Conference on*, pp. 588-597, 2016.

Xing-kun Han is a graduate student at the School of Information Science and Engineering of Xinjiang University. His research interests include image processing and pattern recognition. He is now researching script identification of Central Asia in Xinjiang Laboratory of Multi-language Information Technology. He is a member of IAENG.

Alimjan Aysa is an associate professor at Network and Information Technology Center of Xinjiang University. He also has been working as a researcher in Xinjiang Laboratory of Multi-language Information Technology since 2002. He received PhD degree in computer application from School of Information Science and Engineering, Xinjiang University, China in 2014. His research interests include natural language processing, pattern recognition, and digital signal processing. He has published one book and more than 30 papers. His work has appeared in many Chinese core journals such as Computer Engineering and Application, Chinese Information Processing, and Computer Application etc. Dr. Aysa is a member of Pattern Recognition Professional Committee for China Artificial Intelligence Association (CAAI-PR), IEEE Computer Society (IEEE-CS), and China Computer Federation (CCF).

Hornisa Mamat is an assistant professor (lecturer) at School of Information Science and Engineering, Xinjiang University, China. She received M.S degree from Beijing Institute of Technology in 2012. Her research interests include image processing, pattern recognition, and signal processing. She is a member of Pattern Recognition Committee for China Artificial Intelligence Association (CAAI-PR), and China Computer Federation (CCF).

Kurban Ubul is a professor at School of Information Science and Engineering, Xinjiang University, China. He also has been working as a researcher in Xinjiang Laboratory of Multi-language Information Technology since 2000. His research interests include image processing, pattern recognition, speech signal processing, digital signal processing, and education technology. He has published three books, one book chapter and more than 60 papers, presented at numerous international conferences. Professor Ubul currently serves as editor of Singapore Journal of Scientific Research, Cyber Journals, image and signal processing and Journal of Security and Safety Technology. He has reviewer of IEEE IHMS, IET Biometrics, IJITM, and he served as technical committee member/reviewer of many international conferences such as CCBR2017, IGS2017, CCBR2016, CSA2016, Isca2015, ICDAR2015, ICMIAC2014, WOSSPA2013, ISIEA2012, ICSP2012. He is a committee member of Pattern Recognition Professional Committee for China Artificial Intelligence Association (CAAI-PR). Also, he is a committee member of Pattern Recognition and Machine Intelligence Professional Committee for China Automation Association (CAA-PRMI). He is a member of Institute of Electrical and Electronics Engineers (IEEE), International Association of Pattern Recognition (IAPR), IEEE Computer Society (IEEE-CS), Association for Computing Machinery (ACM), and China Computer Federation (CCF).