SCRIPT IDENTIFICATION BASED ON FUSED TEXTURE FEATURE OF CURVELET TRANSFORM SUB-BANDS

Li SHUN[1], Alimjan AYSA[2], Hornisa MAMAT[3], Yali ZHU[4], Kurban UBUL[5] *

In order to improve the effect of script identification, a script identification method based on fused texture feature of curvelet transform sub-bands was proposed. The texture features of the high-frequency sub-band and the low-frequency sub-band of curvelet transform were extracted, and the statistical features of the image were fused to form 20-dimensional features. SVM, KNN and linear discriminate analysis (LDA) were used to classify these features to complete the experiment. A total of 10,000 document images in 10 kinds of scripts—English, Chinese, Arabic, Russian, Turkish, Mongolian, Kyrgyzstan, Kazakhstan Uyghur, and Tibetan —were classified. More than 99.32% identification rate was obtained by using the SVM classifier. The proposed method needed less dimension features and time in calculation. The experimental result showed that the identification rate of the proposed method was better than that script identification methods based on wavelet transform and dual-tree complex wavelet transform.

Keywords: Script identification, Fused texture feature, Curvelet transform, SVM

1. Introduction

With the increasing communication between countries, optical character recognition (OCR) technology has become more and more common. As the front-end processing technology, script identification technology has become particularly important. In the mass information processing, the script identification technology has become a hot topic of extensive research.

There are three main kinds of script identification methods. One based on statistical characteristic [1, 2], one based on symbol matching [3, 4] and the other methods based on texture feature [5-16]. The first two methods are easily affected by document image tilt, noise and so on. Therefore, these methods are relatively poor robustness. Texture is visual feature that reflects the homogeneity of the image. These features will appear in the document images with different scripts. In recent years, the script identification methods based on texture feature has

[1] M.S., Student, Sch. of Information Science and Engineering, Xinjiang University, Urumqi, China
[2] Prof., Network and Information Center, Xinjiang University, Urumqi, China
[3] Lecture., Sch. of Information Science and Engineering, Xinjiang University, Urumqi, China
[4] Lecture., Sch. of Information Science and Engineering, Xinjiang University, Urumqi, China
[5] Prof., Sch. of Information Science and Engineering, Xinjiang University, Urumqi, China,
*Corresponding author mail: kurbanu@xju.edu.cn

gradually favored. Li [5] proposed a script identification algorithm based on Gaussian derivative filter bank, which could extract the edge and ridge features of script in more directions. The wavelet transform [6] has fast algorithm and a small amount of computation, the multi-resolution decomposition of each level extracted the features of the three directions. Hasimu [7] analyzed the characteristics of special characters, compound characters and certain characters in the three scripts of Uygur, Kazakh, and Kyrgyz. Paper [8] proposed a global deep learning model that combines depth features and mesoscale representations. The authors [9] used convolutional neural networks to identify multiple-script recognition at the element-level discriminant learning method. Singh et al. [10] extracted texture features from the document pages based on the Gray Level Co-occurrence Matrix (GLCM). The paper [11] outcome of the present experiment reveals the usefulness of the Modified log-Gabor filters-based features in recognition of handwritten Indic scripts. Ferrer et al. [12] has proposed a novel method for line-wise script detection based on script character stroke distribution. The stroke distribution is obtained via local patterns. Mijit et al. [13] extracted six texture features from page image that to find the sensitivity of each feature for the document image, it was determined the optimal weights suitable for identification of central Asian multilingual scripts. Han et al. [14, 15] proposed methods based on Nonsubsampled contourlet transform (NSCT), and extracted texture features the local binary patterns and the GLCM features. The paper [16] reviewed the research activities in this field. It highlights the achievements made so far that have important value.

Although script identification technology based on texture feature has been developed for many years and obtained promising results. Many methods were aim at scripts of certain regions and countries and they cannot be applied to more scripts. According to some research results, the single texture features can not fully express the texture information, and the texture feature fusion-based method can make up for this defect [14]. In view of some existing problems of identification methods, this paper proposed a script identification method based on fused texture feature of curvelet transform sub-bands [17, 18]. The features of mean, energy, variance, third-order moment and fourth-order moment were extracted from high frequency sub-band and low frequency sub-band after curvelet transformation, and the statistical features of the image were fused to form 20-dimensional features. A total of 10000 document images contains 10 scripts such as English, Chinese, Arabi, Russian, Turkish, Kazakh, Kyrgyzstan, Mongolian and two minority scripts in China (Uyghur, Tibetan) were used in our experiments. The identification efficiency of SVM, KNN and LDA were analyzed and compared.

2. Curvelet Transform

Curvelet transform development on the basis of ridgelet transform, multiscale ridgelet transform and bandpass filter theory. At a sufficiently small scale, the curve can be viewed as a straight line, and the curve's singularity can be represented by a straight line's singularity. When the image has a singular curve in two-dimension and the curve is quadraticly differentiable, the curvelet can adaptively track the singular curve.

This kind of basis also has directionality. So, we can make curve singularity more sparse representation than wavelet transform. Taking the Cartesian coordinate system $f\ [t_1,\ t_2]$ and $0 \le t_1,\ t_2 \le m$ as input, the discrete form of the curvelet transform is:

$$P^D(j,l,k) = \sum_{0 \le t_1, t_2 < m} f\left[t_1, t_2\right]\overline{\varphi^D_{j,l,k}\left[t_1, t_2\right]} \tag{1}$$

where $\varphi^{\mathrm{D}}_{j,k,l}$ represent the curvelet function, j, k and l represents scale, direction and position parameter respectively.

In the implementation of the two-fast discrete curvelet transform methods proposed by Candes and Donoho, we chose the wrapping-based transform. Analyzed the coefficient of an image (128*128, bmp format) have taken curvelet transform. Obtained the structure coefficient $P\{j\}\{l\}(k_1, k_2)$, j represents the scale, l represents the direction and (k_1, k_2) represents the coordinate in the direction on the scale layer. The image coefficient structure is shown in Table 1.

Table 1

Identification efficiency comparison

level	Scale factor	number	Matrix form			
Coarse	P{1}	1	21*21			
Detail	P{2}	16	18*22	16*22	22*18	22*16
	P{3}	32	34*22	22*34	32*22	22*32
Fine	P{4}	1	128*128			

The image was divided into four scales after it has been curvelet transform. The first layer is called coarse scale layer, which is a matrix of low-frequency coefficients; the second and third layers are called detail scale layer, which composed of medium and high frequency coefficients; the fourth level is called fine scale that consist of high-frequency coefficients. The low-frequency coefficient contains the overview of the image, the high-frequency coefficient reflects the image details and edge features.

3. Proposed Method

Script identification includes the following sections: the document image database establishment, preprocessing, feature extraction and classification.

Firstly, the page document images are obtained by scanning a paper document, like books and newspapers. Then, the standard document images (128*128, bmp format) are acquired by cropping the whole page document image. These standard images of the same size built up the experimental database. These images are pre-processed firstly, and then extracted and saved the feature. These eigenvectors are randomly divided into two parts: a training set and a testing set, which are trained and tested by a classifier.

A. Document image Acquisition and Preprocessing

There are no unified document image libraries for us to study of script identification, so we set up our own experimental database. Scanned different scripts on books and newspapers and saved. The full-page images are cut into the same sized 128*128 dpi images. The experimental database containing 1000 images in each script and a total of 10000 document images are included. The database containing Arabic (Ar), Russian (Ru), Chinese (Ch), Kazakh (Ka), Kyrgyzstan (Ky), Mongolian (Mo), Turkish (Tu), Uyghur (Uy), English (En) and Tibetan (Ti). The selection of scripts includes general cultural species at domestic and abroad, central Asian and Arabic scripts. It has universal applicability. Fig. 1 shows some samples in the experimental database.

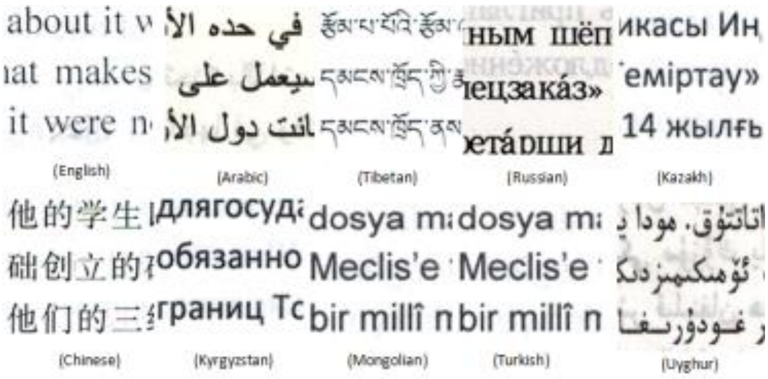
(English)
(Arabic)
(Tibetan)
(Russian)
(Kazakh)
(Chinese)
(Kyrgyzstan)
(Mongolian)
(Turkish)
(Uyghur)

Fig. 1. Some samples in the experimental database

Preprocessing is an important step in the script identification, and it will affect the overall identification performance. Taking into account the database we set up, the paper was thin, and the scanned document image have photocopy on the other side. Therefore, chose grayscale and image binarization as the document image preprocessing.

B. Feature Extraction

The low-frequency coefficients and the high-frequency coefficients which contain rich texture feature information are used for feature parameter for classification. Calculate the mean, energy, variance, third-order distance and fourth-order distance of low-frequency coefficients. Calculate the energy, variance, third-order distance and fourth-order distance of high-frequency coefficients. Obtain a 9-dimensional feature vector.

Mean

$$M_1=\sum_{x=1}^{m}\sum_{y=1}^{n}P_{xy}\{1\}\{1\}/(m\times n) \tag{2}$$

Energy

$$E=\sum_{x=1}^{m}\sum_{y=1}^{n}P_{xy}\{i\}\{1\}/(m\times n) \tag{3}$$

Variance V and third-order distance T is:

$$V=(\sum_{x=1}^{m}\sum_{y=1}^{n}P_{xy}\{i\}\{1\}-M_1)^2 \tag{4}$$

$$T=(\sum_{x=1}^{m}\sum_{y=1}^{n}P_{xy}\{i\}\{1\}-M_1)^3 \tag{5}$$

where, *m* and *n* are the size of cell matrix. According to [15, 16], the fusion of features improves the identification efficiency of a single texture feature. The document image was used for a matrix to calculate statistical feature, which are fused with the transformed features. Calculate the average value of the color image, the mean value after graying, the standard deviation of the image, the image entropy, the local entropy, the local standard deviation, the standard deviation after the histogram equalization processing and the correlation coefficient of the image. Thus, 20-dimensional feature are finally taken as input.

Standard Deviation

$$s=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-x)}^{2} \tag{6}$$

where $x=\frac{1}{n}\sum_{n}^{1}x_i$, $i=1,2,\cdots,n$.

Correlation Coefficients

$$r=\frac{\sum_{m}\sum_{n}(A_{mn}-\overline{A})(B_{mn}-\overline{B})}{\sqrt{(\sum_{m}\sum_{n}(A_{mn}-\overline{A})^2)(\sum_{m}\sum_{n}(B_{mn}-\overline{B})^2)}} \tag{7}$$

where *m* and *n* are gray-scale images of rows and columns, is the mean of the matrix and is the mean of the matrix B.

Wavelet transform (WT) is a multiresolution representation methods, it has a wide range of applications in the extraction of texture features. The script identification based on wavelet transform decomposes the image into three levels of wavelet and obtain nine detail subgraphs. Calculate the wavelet energy feature and the proportion of all energy in the same scale of wavelet energy. Get an 18-dimensional eigenvector. The dual-tree complex wavelet transform (DTCWT) has approximate translational invariance and well direction selectivity in image processing. The method based on dual-tree complex wavelet transform decomposes the image into two levels of dual-tree complex wavelet and obtain six detail high frequency detail subgraphs. Calculate the dual-tree complex wavelet energy feature and the proportion of all energy in the same scale of dual-tree complex wavelet energy. Get a 12-dimensional eigenvector.

The average energy of a K*K image is defined as follows:

$$E = \frac{1}{K^2} \sum_{m=0}^{K-1} \sum_{n=0}^{K-1} f^2(m,n) \tag{8}$$

After the multiscale decomposition of the image, the average wavelet energy of each detail subgraph is defined as:

$$ED^i_{j,m,n} = \frac{1}{K^2} \sum_{m=0}^{K-1} \sum_{n=0}^{K-1} | D^i_{j,m,n} |^2 \tag{9}$$

where $D^i_{j,m,n}$ is the high frequency detail subgraph, *j* is the decomposition scale, *i*=1,2,3.

C. Classifier Selection

The principle of LDA is to project data points into a lower dimensional space and the projected data points are differentiated by categories. Points of the same category are closer together in the projected space. Different types of points are farther away after projection. For a classification problem of K-classification, there is a linear function:

$$y_k(x) = w_k^T x + w_{k0} \tag{10}$$

For all *j*, there is $y_k > y_j$, *x* belongs to category *k*. For each category, there is a formula to calculate a score. Among the scores obtained from all formulas, the largest is the category. KNN classifier measure the distance between different features values to complete classification. In the feature space, most of the k-similar samples of a sample belong to a certain class; the sample also belongs to this class. The main parameter setting of KNN is the choice of k value and distance function. In this paper, the best value of k is set to 3 and “cityblock” distance is selected. SVM map the sample space to a high-dimensional feature space through a nonlinear mapping. The problem of nonlinear separability in the original sample space transformed into a linear separable in the feature space. In this paper, SVM_GUI_3.1 toolbox was used to classify document images. The number of cross-validations is 5. And the radial basis function shown in equation (11) is chosen as the kernel function of SVM.

$$K(x_i, x_j) = \exp(\frac{-\left\| x_i - x_j \right\|^2}{2\sigma^2}) \tag{11}$$

4. Experimental Results and Analysis

During the experiment, features of the four methods were extracted from the database. For all experiments, 10% to 80% of the image features were randomly selected for training set and the rest were used for testing. The average of 10 experimental results was used as the identification result. The experimental platform for this paper is an Inter (R) Core i5-6500 @ 3.20 GHz processor with MATLAB 2016a on a Windows 7 64-bit system with 4GB of memory.

The identification efficiency is evaluated by the feature dimension D and the extraction feature time T. The identification effect evaluates by the recall rate R.

$$R = \frac{n_c}{n} \tag{12}$$

where n_c is the number of samples correctly identified in the script, and n is the number of test sets in the script.

A. Comparative Experiments of Different Methods

Script identification experiments were carried on using curvelet transform with KNN, LDA and SVM respectively, and experimental results were indicated as the following Fig. 2. It can be seen from the Fig. 2 that the identification rates obtained by the SVM were higher than KNN and LDA.

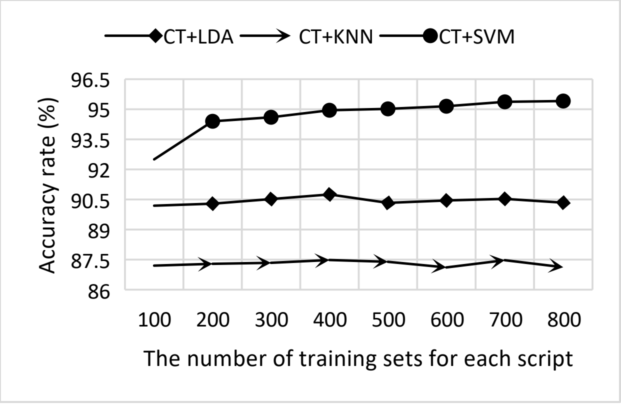

Fig. 2. Script identification results based on curvelet transform (CT)

When the number of training set from each script image was 500, the highest accuracy rate was 95.41% which classified by SVM. The highest identification rates obtained by KNN and LDA were 87.48% and 90.75% respectively. When the number of training set increased, the average identification rate obtained by LDA increased steadily.

In order to prove the effectiveness of the proposed method, it was compared with two classic methods-the script identification based on wavelet transform (WT) and the method based on dual-tree complex wavelet transform (DTCWT). The identification results as shown in Fig. 3 and Fig. 4 separately.

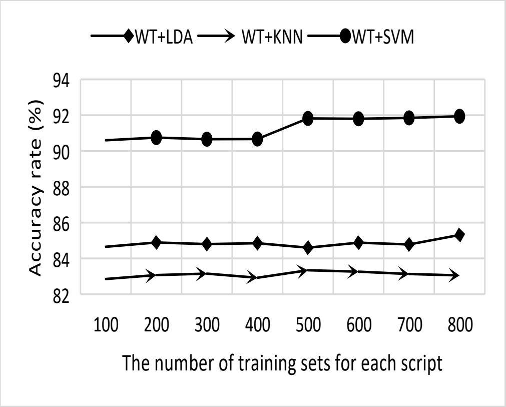

Fig. 3. WT based identification method

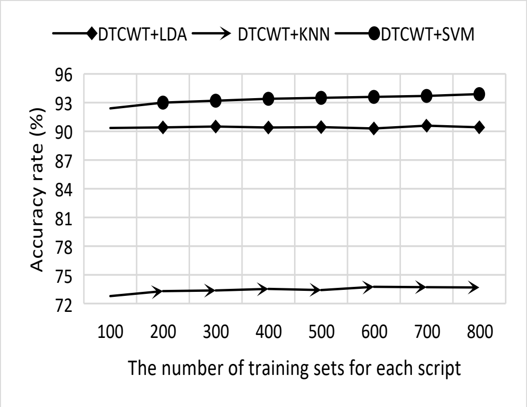

Fig. 4. DTCWT based identification method

The maximum average recall rates of the three kinds of classifier based on wavelet transform were 84.88%, 83.34% and 91.94% respectively. The result obtained by the LDA and KNN classifier is stable at around 84.50% and 83.00%. The maximal recall rates of the script identification method based on dual-tree complex wavelet transform used the three classifiers were 90.59%, 73.75% and 93.89% respectively.

It can be seen from the comparison of Fig. 2-Fig. 4 that the method based on curvelet transform proposed in this paper extracted 9-dimensional texture features for classification and identification and extracted a smaller number of feature dimensions than the other two classical methods. The results under three kinds of classifier were better than those methods based on wavelet transform and dual-tree complex wavelet transform. It was shown that the proposed method can extract texture feature better and improve the identification effect of scripts. The highest accuracy rate was 95.41% based on curvelet transform, it did not meet the high precision requirements for script identification. Therefore, the statistical features of the image are fused to improve the recognition rate. Fig. 5 shows the script identification result based on fused texture feature of curvelet transform sub-bands. Similarly, when used the SVM classified, the identification results were higher than use KNN and LDA. It also proves the superiority of SVM in processing image classification. As the curve shown, the training set increased, the curve was stable above 99% when using SVM. The highest identification result was 3.97% higher than the method based on curvelet transform. When the training set of each script was 100, which are 10% of the total number of database, the identification result reached to 98.77% by using SVM. The result has satisfied the accuracy requirement of text categorization. It is shown that the generalization ability of the proposed method is strong and few training set can achieve the goal of accurate classification.

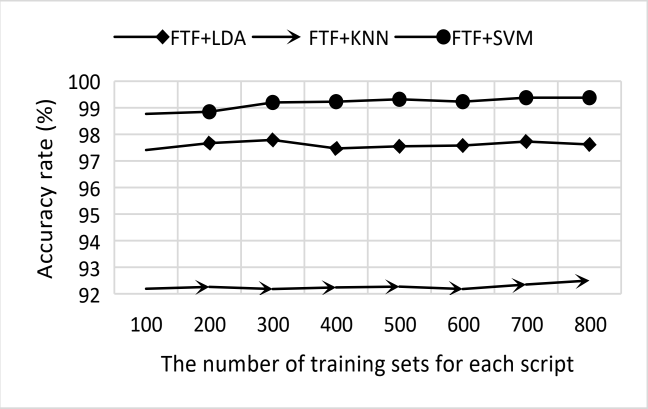

Fig. 5. Script identification based on fused texture feature of curvelet transform (WCT)

It was compared the highest identification results obtained by the above three methods. When using the same classifier, the method presented in this paper has much higher identification results than the other two classic methods. The multi-scale analysis method is a multi-resolution, band-pass and directional function, which can better extract the texture feature and to improve the identification effect. It was compared the identification efficiency of the three methods, indicated in Table 2, where D is feature dimension and T is the feature extraction time.

Table 2

Identification efficiency comparison

Method	D	T/s
WT	18	0.044
DTWCT	12	0.055
FCT	20	0.075

Compared with the methods based on WT and TCWT, the proposed method increased the feature extraction time 0.031s and 0.020s respectively. But it’s maximum identification performance was improved 7.44% and 5.49% respectively. It makes short time sacrifice for better identification accuracy.

B. Error Classification Analysis

The proposed method using SVM obtained the average recall rate of 99.38% when each script training sample was 500. The specific identification effect of the proposed method on each script is analyzed, and it was shown in Table 3.

Table 3

The identification results statistics

Script	**Ar**	**Ru**	**Ti**	**Ch**	**Uy**	**En**	**Mo**	**Ky**	**Tu**	**Ka**
Ar	500									
Ru		500								
Ti			497	1	2					
Ch				494						6
Uy					500					
En						500				
Mo				2			494			4
Ky	1							499		
Tu									500	
Ka				9			6			485
R(%)	100.00	100.00	99.40	98.80	100.00	100.00	98.80	99.80	100.00	97.00

For scripts with large differences in texture structure, the proposed method achieved error-free identification. For similarity scripts, like Kazakh and Kyrgyz belong to the Copchak (Kincha) language of the Altai and Turkic branches. These scripts are similar to 90%. Therefore, the identification accuracy of these similarity scripts was lower. But the lowest script accuracy rate was 98.80%.

C. Comparison with Previous Methods

In order to illustrate the effectiveness of the proposed method, this paper summarized the identification result of the previous methods for script identification as shown in Table 4. As shown in the Table 4, the Gaussian derivative filter bank designed by Tong Li et al. [5] obtained a 98.61% identification rate under the SVM classifier.

Table 4

The comparison results

Author	Method	Classifier	Train set	Test set	T_R (%)
Tong Li [5]	Gaussian derivative filter bank	SVM	1000	2000	98.61
M. Hasimu [7]	Unique character	N/A	70	70	96.67

MA. Ferrer [12]	LBP	SVM	250	500	95.41
Buvajar Mijit [13]	Weighted fused texture	LDA	100	100	95.69
Xing-kun Han [14]	NSCT	KNN	4000	3000	98.91
Proposed Method	CT	SVM	5000	5000	95.41
	FCT		5000	5000	99.32

Hasimu [7] analyzed the unique characters, compound characters and the special features of some characters in certain language context, obtained 96.17% accuracy with 70 words more. Ferrer et al. [12] used LBP to extract the texture features and used the SVM classifier to train and test, with a 95.41% accuracy rate. Mijit [13] extracted fusion futures and classified by LDA, and get 95.69% of recognition rate. Han [14] fused texture feature of NSCT sub-bands and used the KNN got 98.91% accuracy rate. In the case of 5000 training samples and 5000 test samples achieved 99.38% accuracy rate. The proposed method based on fused texture feature of curvelet transform sub-bands got better results than these above methods.

5. Conclusion

The script identification method based on fused texture feature of curvelet transform sub-bands was proposed in this paper. The standard document image database contains 10 scripts and 10000 document images were built. These scripts were selected to consider the similarity and special structure of characters, that is , it include similar shaped scripts, such as Latin (English, Turkish), Cyrllic (Russian, Kazakh, Kyrgyzstan), Arabic (Arabic, Uyghur), and some special scripts (Mongolian, Tibetan). Then the features of mean, energy, variance, and third-order moment were extracted separately after using the texture features of the high-frequency sub-band and the low-frequency sub-band of curvelet transform, and the statistical features of the image were fused to form a 20-dimensional feature vector. Three types of classifier: SVM, KNN and LDA were used to classify the texture feature to complete the experiment respectively. It was obtained 99.32% of average accuracy using SVM when the training set is 500 image in each script image. It needs less dimension features and time in calculation. The experimental result showed that it is robust to font size, format, noise, stroke breakage and so on. Also, the identification rate of the proposed method was better than other script identification methods.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant (No. 61862061, 61563052, 61363064), The Funds for Creative Groups of Higher Educational Research Plan in Xinjiang Uyghur Autonomous, China under Grant (No. XJEDU2017T002), and 2018th Scientific Research Initiate Program of Doctors of Xinjiang University under Grant No. 24470.

REFERENCES

[1]. *J. Jin, R. Cui, X. Cui*, "Script identification for document images between Chinese and Korean language based on wavelet statistic features at level of text row", Journal of Yanbian University (Natural Science), **vol. 39**, no. 4, 2013, pp. 277-280.

[2]. *M. M. Goswami, K. Mitra*, "Classification of Printed Gujarati Characters Using Low-Level Stroke Features", ACM, **vol. 15**, no. 4, 2016, pp. 25-29.

[3]. *G. Wang, Y. Jin, L. Liu, R. Chu*, "East Asian Script Identification Based on Multi-feature", Computer Science, **vol. 40**, no. 1, 2013, pp. 273-276.

[4]. *S. Singh, A. Kumar, D. K. Shaw, D. Ghosh*, "Script separation in machine printed bilingual (Devnagari and Gurumukhi) documents using morphological approach", Communications, 2014, pp. 1-5, doi: 10.1109/NCC.2014.6811361.

[5]. *L. Tong, L. Zhou, X. Ping, S. Xu*, "Script identification based on gaussian derivative filter bank", Journal of Data Acquisition and Processing, **vol. 29**, no. 5, 2014, pp. 713-719.

[6]. *P. Hill, A. Achim, ME. Al-Mualla, D. Bull*. "Contrast Sensitivity of the Wavelet, Dual Tree Complex Wavelet, Curvelet and Steerable Pyramid Transforms", IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society, **vol. 25**, no. 6, 2016, pp. 2739-2751.

[7]. *M. Hasimu, W. Silamu, W. Mushajiang, N. Youliwai*. "Unique character based statistical language identification for Uyghur, Kazak and Kyrgyz", Journal of Chinese Information Processing, **vol. 29**, no. 2, 2015, pp. 111-117.

[8]. *B. Shi, X. Bai, C. Yao*, "Script identification in the wild via discriminative convolutional neural network", Pattern Recognition, **vol. 52**, no. 282, 2016, pp. 448-458.

[9]. *M. Mehri, P. Gomez-Krämer, P. Héroux, et al.*, "A texture-based pixel labeling approach for historical books", Pattern Analysis & Applications, **vol. 20**, no. 2, 2017, pp. 325-364.

[10]. *P.K. Singh, SK. Dalal, R. Sarkar, M. Nasipuri*, "Page-level script identification from multi-script handwritten documents", International Conference on Computer, 2015, pp. 1-6.

[11]. *PK. Singh, I. Chatterjee, R. Sarkar*, "Page-level handwritten script identification using modified log-Gabor filter-based features", IEEE International Conference on Recent Trends in Information Systems, 2015, pp. 225-230.

[12]. *MA. Ferrer, A. Morales, U. Pal*, "LBP Based Line-Wise Script Identification", International Conference on Document Analysis & Recognition, 2013, pp. 369-373.

[13]. *B. Mijit, K. Ubul, N. Yadikar, T. Yibulayin, A. Aysa*, "Weighted Fusion of Texture Features based Central Asian Multi-scripts Identification", Computer Engineering and Applications, **vol. 53**, no. 20, 2017, pp. 187-194.

[14]. *X. Han, A. Aysa, N. Yadikar, Y. Zhu, K. Ubul*, "Script identification of Central Asian based on fusioned texture feature of NSCT sub-bands", Computer Engineering and Applications, 2017, 9.

[15]. *X. Han, A. Aysa, H. Mamt, K. Ubul*, "Script identification of central asian printed document images based on nonsubsampled contourlet transform", Engineering Letters, **vol. 25**, no. 4, 2017, pp. 389-395.

[16]. *K. Ubul, G. Tursun, A. Aysa, D. Impedovo, G. Pirlo, T. Yibulayin*, "Script Identification of Multi-Script Documents: A Survey", IEEE Access, **vol. 5**, no. 99, 2017, pp. 6546-6559.

[17]. *J. L. Starck, E. J. Candes, D. L. Donoho*, "The curvelet transform for image denoising", IEEE Transactions on Image Processing, **vol. 11**, no. 6, 2002, pp. 670-84.

[18]. *J. Ma, G. Plonka*, "The Curvelet Transform", IEEE Signal Processing Magazine, **vol. 22**, no. 2, 2010, pp. 118-133.