Uyghur Printed Document Image Retrieval based on Hu Invariant Moment Features

Kurban Ubul
School of information science and engineering
Xinjiang University
Urumqi, China, 830046
kurbanu@xju.edu.cn

Nurbiya Yadikar
School of information science and engineering
Xinjiang University
Urumqi, China, 830046
nurbiya611@163.com

Ayxemgul Amat
School of information science and engineering
Xinjiang University
Urumqi, China, 830046
569458581@qq.com

Yasen Aizezi
Xinjiang Police Institute
Urumqi, China, 830046
yasinjan@sina.com.cn

Yunus Aysa
School of information science and engineering
Xinjiang University,
Urumqi, China, 830046
yunus@xju.edu.cn

Tuergin Yibulayin*
School of information science and engineering
Xinjiang University
Urumqi, China, 830046
Corresponding author mail: turgun@xju.edu.cn

Abstract—Document image retrieval technology is extensively studied, but there is no report about Uyghur document image retrieval. Hu invariant moment features based document image retrieval scheme is proposed for Uyghur document images in this paper. Firstly, seven types of invariant moment features are extracted from Uyghur document images after obtaining image edge information using Canny edge operator. Then, the features are matched using Euclidean distance classifier and Feature distance classifier between query image with the target images. It is obtained the search results after ordering candidate images according to their similarity. Two types of experiments which are different in query image size are conducted using 1948 Uygur printed document images. The experimental results show that, the highest document image retrieval efficiency (to be get the matching rate of 100%) is obtained here when using the hole page document image is to be set as query image, and the matching rate is declined when sub images are selected as queries, the more the number of sub images, the lower the matching efficiency, and the retrieval efficiency is reached at minimum level when using one sixteenth of the document as query in our experiment. The experimental results indicated that Hu invariant moment features can effectively describe the nature of the Uyghur document images.

Keywords- Uyghur; Document image retrieval; Hu Invariant Moments Features; feature matching

I. INTRODUCTION

Document image retrieval is a new developed research method which is based on the given query image to retrieve similar document image from huge number of document image databases by using matching and retrieval techniques. A detailed research of printed document image retrieval technique can be started in 1994 year. Two years later, research of handwritten document image retrieval has carried out by R. Manmatha et al. [1]. J.J. Hull [2] first addressed the content based matching and retrieval algorithm. But, it is only suitable for small database. C.L. Tan et al. [3] proposed text retrieval method for English and Chinese documents. In it, documents are segmented into characters, vertical traverse density (VTD) and horizontal traverse density (HTD) features are extracted, and similarity between documents is measured by calculating the dot product of document vectors. S. Ediz et al. [4] proposed retrieval system for historical Ottoman documents, different shape features are extracted, and partial symbol wise matching is used to retrieve word images. H. Dewen et al. [5] make retrieval researches on Japanese documents. In it, hierarchical matching tree algorithm is proposed; matching model and texture character strings are described for indexing. Rath, T. M. et al [6] presented a keyword retrieval system for historical documents, and the DTW word image matching algorithm was evaluated for matching image words. Experimental results show that the DTW matching algorithm is much suitable for handwritten or historical documents to retrieval information than other matching methods. G. Kumar et al. [7] proposed a script independent keyword spotting system for multilingual handwritten documents. Local character level score and global word level hypothesis scores are calculated, and then by using a Bayesian logistic regression classifier to distinguish keywords and non keywords. So far, many new algorithms and methods are used for document image retrieval.

Hu invariant moments is a one of the common algorithm in region extraction methods. At present, although studies on Hu moment invariants based image retrieval [8] is also increasing, ant it is used other aspects such as the transportation device identification [9], dental X-ray film image recognition [10], and other fields have gotten a certain amount of research results, but most of them are based on characters and various scenes images as research subjects, researches on document image retrieval are basically don't have, Uyghur document image retrieval even more so. So Hu invariant moments based document image retrieval have great significance, especially for finding and effective use of the Uyghur documents have significant role and practical value.

The Uyghur are a Turkic-speaking ethnic group inhabiting Eastern and Central Asia. Today, Uyghurs mainly live in the Xinjiang Uyghur Autonomous Region (hereafter: Xinjiang) in China. Arabic based Uyghur script is an official writing system in Xinjiang, while Cyrillic based Uyghur script is still used by Uyghurs in former Soviet Union Republics and Latinbased Uyghur script are also in use. The Arabic-based Uyghur script (hereafter: Uyghur) used widely in Xinjiang area is studied in this paper. Uyghur character is composed of 32 letters including 8 vowel letters and 24 consonant letters, besides 4 kinds of different forms for each character. Thus, 32 letters become more than 120 character styles [11].

II. Document image feature extraction

Hu invariant moment is an image region based moment feature, is a seven moments of the image, and is relatively stable image shape features. To obtain shape information of an image, must do edge detection on the image [8]. In this paper, based on the use of canny operator it has been calculated the edge of the document image and fulfilled the extraction of Hu moment invariant feature. Hu invariant moment features are after calculating the image edges. The two steps are introduced briefly here.

A. Calculating the image Edge

Calculating step of the image edges include as following four steps:

1) Use Gaussian filter to smooth the image. Two-dimensional Gaussian function is:

$$G(x,y)=\frac{1}{2\pi\delta^2}\exp\left(-\frac{\left(x^2+y^2\right)}{2\delta^2}\right) \qquad (1)$$

In a direction n, the first order directional derivative of is:

$$G_n=\frac{\partial G}{\partial n}=\mathrm{v} \mid G \qquad (2)$$

$$\mathrm{v}=\begin{bmatrix}\cos\theta\\ \sin\theta\end{bmatrix},\quad \nabla G=\begin{bmatrix}\frac{\partial G}{\partial x}\\ \frac{\partial G}{\partial y}\end{bmatrix} \qquad (3)$$

Where, n is direction vector, ▽is gradient vector. The image f(x,y) and Gn for convolution, Same time changing the direction of n, n is a direction orthogonal to the detected edges when Gn * f(x,y) obtain the maximum value.

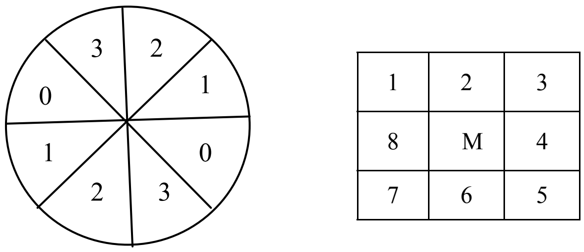
2) Calculate the image gradient (magnitude) and direction by using finite differences of the first order partial derivative.

$$E_X=\frac{\partial G}{\partial x} * f(x,y),\quad E_y=\frac{\partial G}{\partial y} * f(x,y) \qquad (4)$$

$$A(x,y)=\sqrt{E_X{}^2+E_Y{}^2},\quad \theta=Arc\tan\left(\frac{E_X}{E_Y}\right) \qquad (5)$$

Where, A(x, y) is an edge strength of the image at the point of (x, y) , the is direction vector of the image at the point of (x, y).

3) Conduct non maximum suppression to amplitude of the gradient image. In order to determine the edge to be retained the local gradient maximum point's only get global Gradient is not enough, and it could not be more correct to edge detection out, Therefore, using gradient directions to suppression non- maxima. Shown in Figure 1:



1	2	3
8	M	4
7	6	5

Figure 1 Non-maxima suppression

Shown in Figure 1, the variation range of the normal vectors reduced the circumference in the four sectors, and let M is a center pixel in the field of 3*3, edge strength on the each pixel is compared with edge strength of neighborhood's center pixel M and edge strength of two pixels which are along with gradient line respectively. If the edge intensity of M is less than the edge strength of two adjacent pixels, then let M = 0.

4) Use the double threshold algorithm for edge detection and connection. This process mainly includes the edge discrimination and connection, the main principle is as follows: Firstly, it is determined the high threshold and low threshold value, and then according to the intensity of each pixel edge make comparison with the high threshold and low threshold value. If the edge strength of the pixel is greater than the high threshold put the pixel as an edge point, if less than the low threshold is discriminated as non edge point. If the edge strength is greater than the low threshold and less than the high threshold value, to see whether the pixels in the adjacent pixels have the edge points that greater than the high threshold value and point out those edge points. When connecting those edges, based on the non maximum image effect the two thresholds and form the two corresponding threshold image. First connected those edges of the high threshold image, when connected to reach the ends, to find the edge that can be connected from the low threshold image, thus the entire edge of the image points are connected.

B. Calculate the Hu invariant moments

When calculating the Hu invariant moments [12], if the size of the image f(x, y) is M×N (M is length, N is height), then the *p*+*q* th moment is :

$$m_{pq}=\sum_{x=1}^{M}\sum_{y=1}^{N}x^{p}y^{q}f(x,y)\qquad (\pi,\ \theta=0,1,2,...)\quad (6)$$

The central moment defined as:

$$\mu_{pq}=\sum_{x=1}^{M}\sum_{y=1}^{N}(x-\bar{x})^{p}(y-\bar{y})^{q}f(x,y)\qquad (7)$$

Where, $\bar{x}=m_{10}/m_{00}$ and $\bar{y}=m_{01}/m_{00}$; The normalized central moment η_{pq} is defined as $\eta_{pq}=\mu_{pq}/u_{00}^{\ y}$, where $\gamma=(p+q)/2$, p+q=2,3,4,…. Then, the seven Hu invariant moments $\varphi_{1,}\ \varphi_{2,}\ldots\varphi_{7}$ are calculated separately as :

$$\begin{cases}\varphi_1=\eta_{20}+\eta_{02}\\ \varphi_2=(\eta_{20}-\eta_{02})^2+4\eta_{11}^{\ 2}\\ \varphi_3=(\eta_{30}-3\eta_{12})^2+(3\eta_{21}-\eta_{03})^2\\ \varphi_4=(\eta_{30}+\eta_{12})^2+(\eta_{21}+\eta_{03})^2\\ \varphi_5=(\eta_{30}-3\eta_{12})^2(\eta_{30}+\eta_{12})[(\eta_{30}+\eta_{12})^2-3(\eta_{21}+\eta_{03})^2]\\ \quad+(3\eta_{21}-\eta_{03})(\eta_{21}+\eta_{03})[3(\eta_{30}+\eta_{12})^2-(\eta_{21}+\eta_{03})^2]\\ \varphi_6=(\eta_{20}-\eta_{02})[(\eta_{30}+\eta_{12})^2-(\eta_{21}+\eta_{03})^2]\\ \quad+4\eta_{11}(\eta_{30}+\eta_{12})(\eta_{21}+\eta_{03})(3\eta_{21}+\eta_{03})\\ \varphi_7=(3\eta_{21}-\eta_{03})(\eta_{30}+\eta_{12})[(\eta_{30}+\eta_{12})^2-3(\eta_{21}+\eta_{03})^2]\\ \quad+(3\eta_{21}-\eta_{03})(\eta_{21}+\eta_{03})[3(\eta_{30}+\eta_{12})^2-(\eta_{21}+\eta_{03})^2]\end{cases}\qquad (8)$$

Thus, the seven feature moments are calculated. If the target image is A, it is need to calculate A×7 of feature .

III. Document image feature extraction

A. Document image retrieval

When retrieving at first the query document image feature extraction, according to the similarity measure must be the calculated query document image characteristics and the distance between the target document image features. If xi and gi is feature vectors of inquiry image and feature vectors of target image, the similarity measuring methods used in this paper, that is Feature distances classifier and Euclidean distance classifier is as follows:

Feature distances classifier:

$$d(\vec{x}-\vec{g})=\frac{\sum_{i=1}^{n}|x_i-g_i|}{\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}g_i}\qquad (9)$$

Euclidean distance classifier:

$$\delta_E(\vec{x},\vec{g})=[(\vec{x}-\vec{g})^T(\vec{x}-\vec{g})]^{1/2}=[\sum_{i=1}^{n}(x_i-g_i)^2]^{1/2}\quad (10)$$

Above two formulas, respectively query document image and the target document image feature vectors. If the distance between features greater, the smaller their similarity. Then the retrieval result in descending order according to the similarity of size and displayed one by one. The flowchart of retrieval algorithm is shown as the following Figure 2.

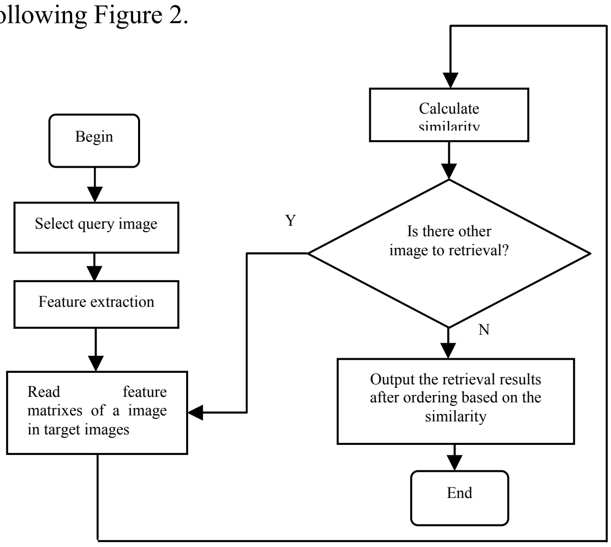
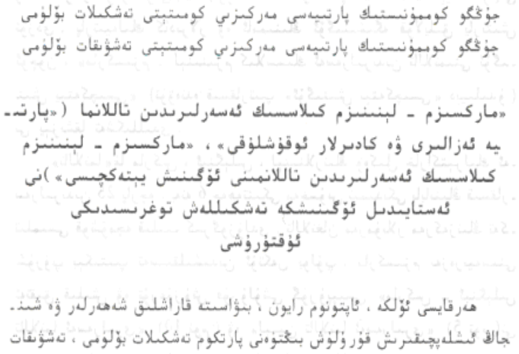

Begin
Select query image
Feature extraction
Read feature matrixes of a image in target images
Calculate similarity
Is there other image to retrieval?
Y
N
Output the retrieval results after ordering based on the similarity
End

Figure 2. Uyghur document image retrieval algorithm processes

B. Experiments and results analysis

In this paper, Uyghur printed document image database that contains 1948 different resolution (100dpi, 200dpi), size is 716 x 1011 pixels of BMP format document image, which mainly contains the text printed document image, and document sample is use A4 printing paper documents and books printed text. A sample of Uyghur printed document image is shown as the following Figure 3.



جۇڭگو كوممۇنىستىك پارتىيەسى مەركىزىي كومىتېتى تەشكىلات بۆلۈمى
جۇڭگو كوممۇنىستىك پارتىيەسى مەركىزىي كومىتېتى تەشۋىقات بۆلۈمى

«ماركسىزم ـ لېنىنىزم كىلاسسىك ئەسەرلىرىدىن تاللانما (»پارتىـ
يە ئەزالىرى ۋە كادىرلار ئوقۇشلۇقى» ، «ماركسىزم ـ لېنىنىزم
كىلاسسىك ئەسەرلىرىدىن تاللانمىنى ئۆگىنىش يېتەكچىسى»)نى
ئەستايىدىل ئۆگىنىشكە تەشكىللەش توغرىسىدىكى
ئۇقتۇرۇش

ھەرقايسى ئۆلكە ، ئاپتونوم رايون ، بىۋاسىتە قاراشلىق شەھەرلەر ۋە شىند.
جاڭ ئىشلەپچىقىرىش قۇرۇلۇش بىڭتۇەنى پارتكوم تەشكىلات بۆلۈمى ، تەشۋىقات

Figure 3. A sample of Uyghur printed document image

In the experiment an entire document images are divided into *m* different don't overlapped sub-images, which are as the retrieval document images and its entire document image as a target document image conducted several experiments, And calculate the average percentage of matching .The results as shown in Table 1 and Table 2.

TABLE 1. Experimental Results With The Feature Distance Classifier

Query condition （query image）	Average matches Percentages （%）	Average retrieval time(Second)
m=1	100%	460
m=2	89.19%	462
m=4	81.51%	461
m=8	55.07%	458

Table 1 shows, when m=1 average percentage matches Highest, when m=2, 4, 8, the average percentage match descending in order, the average retrieval time changes unlikely. Because the calculation Hu invariant moments before the need to calculate the edge of the image, so the average search time is relatively long.

TABLE 2. EXPERIMENTAL RESULTS USING EUCLIDEAN DISTANCE CLASSIFIER

Query condition（query image）	Average matches Percentages（%）	Average retrieval time(Second)
m=1	100%	450
m=2	81.22%	460
m=4	77.96%	465
m=8	54.51%	460

In Table 2, It can be seen the Retrieval efficiency a little higher than Table 1, the average search time is changing unlikely, This indicates that the use of different similarity measures can be earn different retrieval efficiency.

From the above Table1, Table2, it can seen when using the Distance classifier feature retrieval less efficient than using the Euclidean distance classifier retrieval efficiency point, This table show that the Euclidean distance classifier is Uyghur document image retrieval is a better measure of similarity. From figure 4, it is indicated that with the increase of the number m in the efficiency of the image retrieval, correspondingly reduced. This is because as the more the m, between the query document image and the target document images Feature distance is greater. So their similarity is smaller.

IV. CONCLUSIONS

In this paper, a document image retrieval method based on Hu invariant moment features is proposed for Uyghur document images. Uyghur document image database is created after scanning Uyghur text documents with .bmp format firstly. Then, Uyghur document image edge information is obtained by Canny edge operator, and seven types of invariant moment features are extracted. Finally, the features are matched using Euclidean distance classifier and Feature distance classifier between query image with the target images, and search results is obtained after ordering candidate images according to their similarity. Two types of experiments which are different in query image size are conducted using 1948 Uygur printed document images. The experimental results indicate that, the highest document image retrieval efficiency (to be get the matching rate of 100%) in this work is acquired by using the hole page document image is to be set as query image, and the matching efficiency is decreased when sub images are selected as queries, that is, the more the number of sub images, the lower the matching efficiency, and it is reached at minimum level when using one sixteenth of the document as query in our experiment. The average searching time is nearly equal in the different kind experiments. The experimental results indicated that Hu invariant moment features can effectively describe the nature of the Uyghur document images.

As the fist attempt for Uyghur document image retrieval, the Hu invariant moment features shows its strong efficiency. But, the time efficiency is still need to be improved. In our future work, we will try to extract more efficient features for Uyghur document image, and will use more effective matching algorithms to improve the searching efficiency of Uyghur document images.

ACKNOWLEDGMENT

This work is supported by Special Training Plan Project of Xinjiang Uyghur Autonomous Region’s Minority Science and Technological Talents (No. 201323121), College Scientific Research Plan Project of Xinjiang Uyghur Autonomous Region (No. XJEDU2013I11, XJEDU2013I34), the Open Projects Program of National Laboratory of Pattern Recognition (No. 201306321) and the Open Project of Xinjiang Laboratory of Multi-language Information Technology (No. 049807).

REFERENCES

[1] R. Manmatha, C. Han, E.M. Riseman, Word spotting: a new approach to indexing handwriting, Proc. CVPR. 1996: 631-637.

[2] J. J. Hull, Document Image Matching and Retrieval with Multiple Distortion-Invariant Descrip-tors, Document Analysis Systems. 1995:379–396.

[3] C.L. Tan, W.H. Huang, Z.H. Yu, Imaged document text Retrieval without OCR, IEEE Trans on Pattern Analysis and Machine Intelligence. 2002, 24(6) ; 838 – 844.

[4] E. Saykol, A.K.Sinop, U. Güdükbay, Ö. Ulusoy, A.E.Çetin, Content based retrieval of historical ottoman documents stored as textual images, IEEE Trans, Image Process. 2004, 13(3): 314–325.

[5] H.Dewen, W.X.Chang, L.Jiang, A Content based Retrieval Algorithm for Document Image Database, Multimedia Technology (ICMT), 2010 International Conference. 2010: 1 – 5.

[6] T.M.Rath, R. Manmatha, Word spotting for historical documents, IJDAR, 2007 (9): 139-152.

[7] G. Kumar and V. Govindaraju, A Bayesian Approach to Script Independent Multilingual Key-word Spotting, 14th International Conference on Frontiers in Handwriting Recognition. 2014: 357-362.

[8] Qingyong Li,WeitaoLu.Keypoint Based Moment Invariants Descriptor for Ground-based Cloud Image Retrieval. IEEE International Symposium on Industrial Electronics (ISlE 2009) Seoul Olympic Parktel, Seoul, Korea July 5-8, 2009.

[9] QianTian,Tengfeizhong,HongLi. A new method for vehicle detection using mexicanhat wavelet and moment invariants. IEEE Workshop on Signal Processing Systems, 2013

[10] Pattanachai N, Covavisaruch N, Sinthanayothin C. Tooth recognition in dental radiographs via Hu's moment invariants[C]//Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2012 9th International Conference on. IEEE, 2012: 1-4.

[11] Kurban Ubul, Andy Adler and Mamatjan Yasin. Multi-Stage Based Feature Extraction Meth-ods for Uyghur Handwriting Based Writer Identification, Genetic Algorithms in Applications, Dr. Rustem Popa (Ed.), 2012:245-262.

[12] GuoqiangShen, Lan chi Jiang, Guoxuan Zhang. An Image Retrieval Algorithm Based on Color Segment and shape moment invariants. IEEE Second International Symposium on Computational Intelligence and Design, 2009.517-521.