基于灰度直方图与改进Hu不变矩的维吾尔文关键词图像二次检索*

宋志平[1]，朱亚俐[1]，徐学斌[1]，吾尔尼沙·买买提[1]，库尔班·吾布力[1,2†]

(1. 新疆大学 信息科学与工程学院，新疆 乌鲁木齐 830017；2. 新疆多语种信息技术重点实验室，新疆 乌鲁木齐 830017)

摘　要： 维吾尔文文字具有粘连性大、结构不封闭等特点，这给维吾尔文关键词图像检索造成了极大的困难. 为提高维吾尔文文档图像检索效率，提出一种基于灰度直方图与改进Hu不变矩的关键词图像二次检索算法，该算法对单词图像进行两次检索：粗略检索和二次检索. 在粗略检索阶段，对切分后的单词图像提取灰度直方图特征并对单词数据库进行粗略匹配，在保证召回率的情况下，过滤掉部分无关单词图像形成候选单词库. 在粗略匹配的基础上进行精确匹配，使用改进的Hu不变矩对关键词图像的轮廓特征进行描述，该方法在Hu不变矩中将离心率、区域矩和结构矩统一，可以有效地描述图像的轮廓信息. 在包含115张纯文本维吾尔文文档图像数据库上进行实验，其检索准确率平均值为78.36%，召回率平均值为81.68%.

关键词： 维吾尔文；灰度直方图；Hu不变矩；粗略匹配；二次检索

DOI： 10.13568/j.cnki.651094.651316. 2021.04.10.0003

中图分类号： TP391　**文献标识码：** A　**文章编号：** 2096-7675(2022)03-0323-08

引文格式： 宋志平, 朱亚俐, 徐学斌, 吾尔尼沙·买买提, 库尔班·吾布力. 基于灰度直方图与改进Hu不变矩的维吾尔文关键词图像二次检索[J]. 新疆大学学报(自然科学版)(中英文), 2022, 39(3): 323-330.

英文引文格式： SONG Zhiping, ZHU Yali, XU Xuebin, Wuernisha Maimaiti, Kuerban Wubuli. Uygur keyword image secondary retrieval based on gray histogram and improved Hu invariant moment[J]. Journal of Xinjiang University(Natural Science Edition in Chinese and English), 2022, 39(3): 323-330.

Uygur Keyword Image Secondary Retrieval Based on Gray Histogram and Improved Hu Invariant Moment

SONG Zhiping[1], ZHU Yali[1], XU Xuebin[1], Wuernisha Maimaiti[1], Kuerban Wubuli[1,2]

(*1. School of Information Science and Engineering, Xinjiang University, Urumqi Xinjiang 830017, China;*
2. Xinjiang Multilingual Information Technology Key Laboratory, Urumqi Xinjiang 830017, China)

Abstract： Uygur characters have the characteristics of large adhesion and non-closed structure, which makes it very difficult for Uygur keyword image retrieval. In order to improve the efficiency of Uygur document image retrieval, a keyword image secondary retrieval algorithm based on gray histogram and improved Hu invariant moment is proposed. The algorithm retrieves word images twice: rough retrieval and secondary retrieval. In the rough retrieval stage, the gray histogram feature is extracted from the segmented word image and the word database is roughly matched, under the condition of ensuring the recall rate some irrelevant word images are filtered out to form a candidate word library. Accurate matching is carried out on the basis of rough matching and the improved Hu invariant moment is used to describe the contour features of keyword images, this method unifies the eccentricity regional moment and structural moment in Hu invariant moment, which can effectively describe the boundary information of the image. The experiment is carried out on the image database containing 115 plain text Uygur documents. The average retrieval accuracy is 78.36% and the average recall is 81.68%.

Key words： Uygur; gray histogram; Hu invariant moment; rough matching; secondary search

* **收稿日期：** 2021-04-10

基金项目： 国家自然科学基金重点项目（61862061；61563052；61363064）；新疆维吾尔自治区科技厅青年基金项目（2021D01C119）.

作者简介： 宋志平（1994-），男，硕士生，从事数字图像处理的研究，E-mail: 24526710730@qq.com.

† **通讯作者：** 库尔班·吾布力（1974-），男，博士，教授，CCF高级会员（30804S），主要从事数字图像处理与模式识别的研究，E-mail: kurbanu@xju.edu.cn.

0 引 言

随着文档图像规模越来越大，对于英文和中文的光学字符识别[1]（Optical Character Recognition, OCR）技术已经十分成熟. 在OCR识别技术的带动下，人们对于少数民族文字研究也越来越广泛，其中印刷体维吾尔文文档图像检索的研究可以加快维吾尔文数字图书馆的建设与发展，对促进我国少数民族文化的发展具有极其重要的意义.

在文档图像领域中Manmatha等[2−3]第一次提出一种基于词图像匹配的关键词检索算法，在对手稿图像进行词图像分割的基础上实现了手稿图像的检索，在接下来的研究工作中，该团队提取了手写体单词图像的多维轮廓特征，并用动态时间扭曲（Dynamic Time Warping, DTW）算法对单词图片进行检索. Rothfeder等[4]在检索中对样本图像提取Harris角点序列，对查询词也提取同样的特征，度量值设置为像素灰度差的平方和（Sum of Squared Differences, SSD）的计算值. Vadivukarassi等[5]提出了一种提取关键图像特征的HOG-SIFT算法，最后利用子空间聚类算法从图像数据库中提取目标图像. Niaz等[6]提出了一种基于索引和检索的英语文档索引系统，将含有英文文本的文档图像分成连接词，每个文本用一组特征表示，分别提取图像的DCT和DWT特征，然后利用欧氏距离进行关键词检索. 黄祥琳等[7]提出一种不经过ORC识别情况下，将中文文档图像分割成单个中文字符图片，再对中文字符图像的笔画特征数据进行提取，最后利用WMHD进行关键词检索. 魏宏喜等[8]建立了基于词定位法的蒙古文图像检索系统，系统中图像通过文字轮廓、投影特征和笔画的交集来表达，将视觉特征集成到BOVW[9]模型中并用于检索任务.

目前关于维吾尔文文档图像的研究进展迅速，例如周文杰[10]用形态学梯度算法对维吾尔文文档图像进行单词切分，并通过切分后单词图像的LBP特征实现检索. 李静静等[11]提出层级匹配方法实现对关键词进行精确检索. 本文设计了一种由粗到细层次匹配的维吾尔文关键词图像二次检索框架，该方法主要包括4部分：文档图像采集与预处理、单词切分、粗略检索、二次检索，具体框架如图1所示. 首先，对预处理后的印刷体文档图像进行单词切分生成单词数据集，并使用灰度直方图特征对单词图像进行粗略检测，过滤掉无需检索的单词. 其次，在粗略检索的基础上，使用Hu不变矩特征对粗略匹配检索回的候选单词图像库进行二次精确检索.

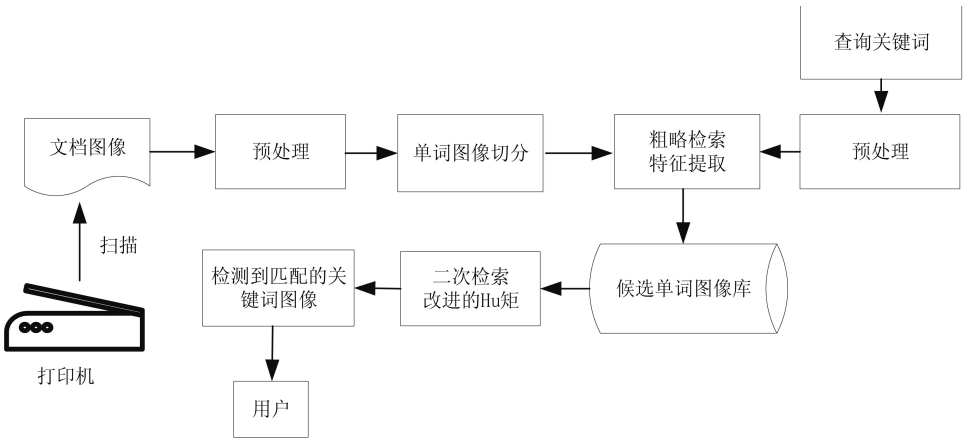


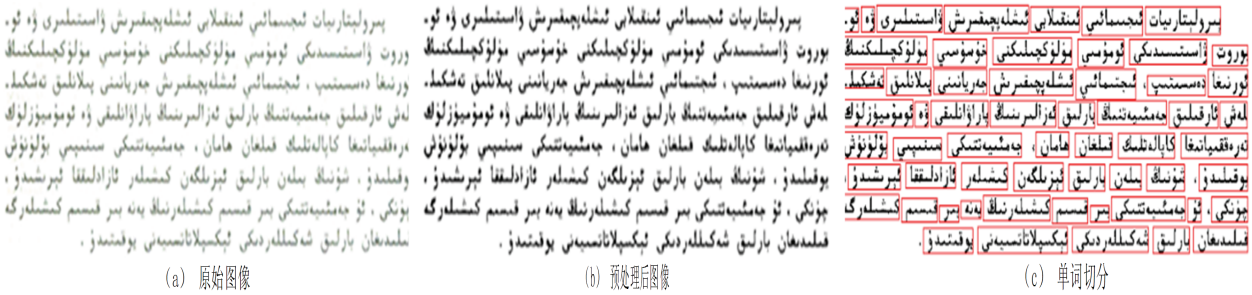
图 1 基于关键词的文档图像系统检索框架

1 基于印刷体关键词检索研究

1.1 文档图像采集与预处理

本文通过扫描维吾尔文版《马列主义经典著作选编》书籍，构建了1 000张纯文本文档图像数据库，随机抽取了115张文档图像进行试验. 在将纸质书籍文本扫描成电子文档图像的过程中，由于噪声、设备及其它因素的干扰，降低了文档图像的质量，从而影响了图像特征有效信息的提取. 因此本文通过图像预处理方法来增强图像质量，预处理操作包括：灰度化（加权平均法）、二值化（Otus）、噪声去除（中值滤波）及倾斜校正（Hough变换）. 预处理前后效果如图2（a）、图2（b）所示，可以看出预处理后图形中的字迹相比处理前更清晰，内容结构也比较容易辨识，这说明预处理后的图像要比灰度图像的特征更加明显. 文献[12]采用连体段的特征聚类方式，将图文混合布局的维吾尔文文档图像进行单词切分，由于本文使用的数据集是纯文本版面，因此本文采用周文

杰等[13]提出的基于形态学分析和像素积分投影算法对维吾尔文文档印刷图像的单词进行切分，效果如图2（c）所示.



(a) 原始图像　(b) 预处理后图像　(c) 单词切分

图 2　文档图像预处理

1.2　特征提取

1.2.1　灰度直方图

灰度直方图[14]（gray histogram）是灰度级的函数，它表示图像中处于该灰度级具有的像素数量. 直方图曲线的横轴是图像的像素灰度等级，纵轴表示该灰度级出现的概率，计算公式如下：

$$H(i)=\frac{n_i}{N}, i=0,1,2,\cdots,L-1 \tag{1}$$

其中：i表示灰度级，L表示灰度级种类数，n表示图像中具有灰度级i像素的个数，N表示图像总的像素数.

1.2.2　灰度直方图性质

在图像检索算法中，灰度直方图算法是最简单的算法之一，该算法容易实现，运算速度快，具有旋转、比例和位移不变性，检索结果不会漏掉相似图像[15]. 但该算法也存在缺陷，在图像中各像素的灰度值都是具有二维位置信息的，而直方图只统计某一灰度值像素的多少和其在图像中的比例，对具有同一灰度的像素在图像中的位置信息无法确定. 这会导致不同的图像具有相同直方图，如图3 (a)、图3 (b)、图3 (c)所示.

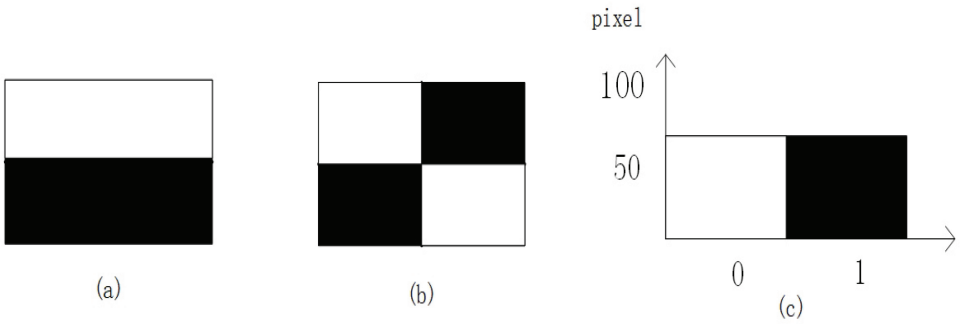


图 3　不同图像具有相同直方图

1.2.3　Hu不变矩

Hu不变矩（invariant momcnts）是Hu[16]在1962年首先提出用来描述图像的形态特征，该特征具有旋转、缩放和平移不变性，因此被广泛用于图像特征提取. 假设图像$f(x,y)$是分段连续的，其$(p+q)$阶笛卡儿系几何矩与几何中心矩μ_{pq}定义为：

$$\begin{cases} m_{pq}=\iint x^p y^q f(x,y)\,\mathrm{d}x\mathrm{d}y, & p,q=0,1,2,L \\ \mu_{pq}=\iint (x-x_0)^p (y-y_0)^q f(x,y)\,\mathrm{d}x\mathrm{d}y, & p,q=0,1,2,L \end{cases} \tag{2}$$

归一化的$(p+q)$阶中心矩η_{pq}定义为：

$$\eta_{pq}=\frac{\mu_{pq}}{\mu_{00}^{\gamma}} \tag{3}$$

其中：$\gamma=(p+q+2)/2$, $p+q=2, 3, \cdots$. 图像的中心距具有平移不变性和缩放不变性的特点. 几何中心矩可用于描述区域的形状.

二维不变矩理论是在笛卡儿坐标系下，通过归一化中心矩η_{pq}从而推导出的7个不变矩，并被证明其对平移、缩放和旋转具有不变性，这为图像的不变矩研究奠定了理论基础，使其广泛的应用在图像处理上，7个二维不变矩的计算公式如下：

$$\begin{cases}\varphi_1=\eta_{20}+\eta_{02}\\ \varphi_2=(\eta_{20}-\eta_{02})^2+4\eta_{22}^2\\ \varphi_3=(\eta_{30}-\eta_{12})^2+(3\eta_{21}-\eta_{03})^2\\ \varphi_4=(\eta_{30}+\eta_{12})^2+(\eta_{21}+\eta_{03})^2\\ \varphi_5=(\eta_{30}-3\eta_{12})(\eta_{30}+\eta_{12})\left[(\eta_{30}+\eta_{12})^2-3(\eta_{21}+\eta_{03})^2\right]\\ \quad+(3\eta_{21}-\eta_{03})(\eta_{21}+\eta_{03})\left[3(\eta_{30}+\eta_{12})^2-(\eta_{21}+\eta_{03})^2\right]\\ \varphi_6=(\eta_{20}-\eta_{02})\left[(\eta_{30}+\eta_{12})^2-(\eta_{21}+\eta_{03})^2\right]+4\eta_{11}(\eta_{30}+\eta_{12})(\eta_{21}+\eta_{03})\\ \varphi_7=(3\eta_{21}-\eta_{03})(\eta_{30}+\eta_{12})\left[(\eta_{30}+\eta_{12})^2-3(\eta_{21}+\eta_{03})^2\right]\\ \quad+(3\eta_{21}-\eta_{30})(\eta_{21}+\eta_{03})\left[3(\eta_{30}+\eta_{12})^2-(\eta_{21}+\eta_{03})^2\right]\end{cases}\tag{4}$$

1.2.4 改进的Hu不变矩

Hu不变矩是基于图像区域且与图像的灰度相关，其计算量大所以Hu矩也称为区域矩. 然而维吾尔文文字构成笔画简单且很多主体笔画相同，其宽度和高度相互之间不等、书写形式与一般字母不同，多由一些弧线和曲线所组成，属于非区域结构. 因此原始的Hu不变矩不能很好地对其边界信息进行有效描述，主要是因为μ_{00}是与区域相关的面积比例因子，然而对于边界像素的计算使得μ_{00}因子失效，不满足边界矩的不变性. 为了得到适用于区域封闭和不封闭结构的统一矩公式，本文利用矩之间的比值来消掉比例因子μ_{00}，从而使不变矩与面积或结构的比例缩放无关仅与几何形状有关. 改进的Hu矩计算公式如下：

$$\Psi_1=\frac{\sqrt{\varphi}_2}{\varphi_1}=\frac{\sqrt{(\eta_{20}-\eta_{02})^+4\eta_{11}}}{\eta_{20}+\eta_{02}}\tag{5}$$

$$\Psi_2=\frac{\varphi_1+\sqrt{\varphi_2}}{\varphi_1-\sqrt{\varphi_2}},\quad \Psi_3=\frac{\sqrt{\varphi_3}}{\sqrt{\varphi_4}},\quad \Psi_4=\frac{\sqrt{\varphi_3}}{\sqrt{\varphi_5}}\tag{6}$$

$$\Psi_5=\frac{\sqrt{\varphi_4}}{\sqrt{\varphi_5}},\quad \Psi_6=\frac{\varphi_6}{\varphi_1\cdot\varphi_3},\quad \Psi_7=\frac{\varphi_6}{\sqrt{\varphi_5}}\tag{7}$$

其中：Ψ_1代表了形状特征的曲率半径，当区域形状为直线特征时曲率半径为1. 当区域形状为圆时曲率半径为0. 为了提高相对矩对形状的描述能力，本文在Hu不变矩中加入离心率特征：

$$\Psi_8=\frac{(m_{20}-m_{02})^2+4m_{11}}{m_{20}+m_{02}}\tag{8}$$

其中：离心率Ψ_8表示形状特征最大轴与最小轴的比值，满足形状特征的集合不变性.

2 实验结果分析

本文实验在Windows10环境下展开，处理器型号为Intel Core I5-8300，运行内存8 GB，具体程序是在python3.7开发环境下编程调试，并借助OpenCV-3.4.2.16开发平台实现. 文档图像数据库源于《马列主义经典著作选编》维吾尔文版书籍，为模拟不同办公环境，采用不同打印机，将纸质文档扫描为文档图像，尺寸为716×1 011，300 dpi. 建立了1 000张文档图像，随机选取了115张文档图像并切分成24 460张单词图像，然后在其中选取10个具有丰富意义的关键词图像作为查询关键词图像，并在数据库中进行检索实验. 检索性能的评价指标有准确率（precision）、召回率（recall）、F值，TP（True Positive）：检索为关键词，实际也为关键词；FP（False Positive）：检索为关键词，实际是非关键词；TN（True Negative）：未被检索到的关键词，实际是非关键词；FN（False Negative）：未被检索到的关键词，实际是关键词，指标计算公式如下：

$$precision=TP/(TP+FP)\times 100\%\tag{9}$$

$$recall = TP/(TP+FN)\times 100\% \tag{10}$$

$$F = precision \times recall \times 2/(precision + recall) \tag{11}$$

2.1　基于灰度直方图的粗略检索实验

本文首先在维吾尔文单词数据集上，评估了灰度直方图算法在维吾尔文文档数据集上的性能，通过调节阈值过滤掉不相关单词图像，在保持召回率的同时尽可能准确地定位目标关键词.

表1展示了灰度直方图特征的检索结果，其检索平均准确率为47.35%，而这些正确的关键词图像占所标注关键词图像总个数的91.33%，F值平均值为61.99%. 由表1可知，第5个关键词准确率与F值在这10个查询关键词中最高，分别为60.52%、69.69%，即表示共检索回76个单词图像，其中目标关键词包含46个. 相反也有准确率较差的关键词，第8个关键词准确率在这10个查询关键词中最差，为36.36%，即共检索到55个单词图像，其中目标关键词图像仅包含20个. 召回率效果最好的是第2、第9两个关键词，为100%. 表示标注关键词均被召回；而召回率最差的是第5个关键词，为82.14%，即从56个标注关键词图像中成功召回46个目标关键词图像.

表 1　基于灰度直方图的粗略匹配结果

	1	2	3	4	5	6	7	8	9	10	平均值
标注关键词	40	69	56	48	56	43	23	23	50	45	-
检索关键词	36	69	50	43	46	40	21	20	50	41	-
共检索单词	88	158	99	90	76	89	46	55	103	75	-
准确率/%	40.90	43.67	50.50	47.77	**60.52**	44.94	45.65	36.36	48.54	54.66	47.35
召回率/%	90.00	**100**	89.28	89.58	82.14	93.02	91.30	86.95	**100**	91.11	91.33
F值/%	56.24	60.79	64.51	62.31	**69.69**	60.60	60.86	51.27	65.35	68.32	61.99

经分析得出，粗匹配阶段准确率普遍偏低，召回率较高. 这是因为该阶段是关键词初步筛选的过程，未对关键词图像提取有效的特征而直接使用像素信息进行匹配，该阶段仅仅对数据库中一些不相关的词图像和切分错误的单词图像进行过滤，以保证尽可能多地召回目标词. 目的在于如果对图像库中所有单词图像均进行特征提取、匹配，明显会增加检索复杂度，做许多无意义的匹配. 因此本文通过阈值的设置，过滤一部分无需匹配的单词构成候选单词集合，并通过中间阈值的设定，在保证召回率的情况下尽可能准确检索到目标关键词.

2.2　基于Hu不变矩的二次检索实验

本节前三组实验是在没有粗略检索的基础上，比较了改进Hu不变矩与原始的Hu不变矩算法在维吾尔文文档图像数据库上检索性能. 表2中数据为使用Hu不变矩算法在维吾尔文文档图像中的匹配结果，其中准确率平均值为39.79%，召回率的平均值为44.21%，F值平均值为36.82%. 其中第5个查询关键词的检索准确率最高，为55.81%. 即共检索到43个单词图像，其中目标关键词包含24个. 准确率最差的是第7个查询关键词，为28.20%，即共检索回39个单词图像，其中正确单词图像为11个. 对于召回率而言，第8个关键词图像召回率最好，为65.21%，表示共标注23个关键词图像，其中有15个目标关键词被召回. 第3个关键词图像召回率最低，为32.14%. 即共标注56个关键词图像，其中召回18个. 表3是采用本文改进后的Hu不变矩算法的检索结果，其准确率平均值为54.22%，召回率平均值为54.27%，F值平均值为53.86%. 其中第6个查询关键词检索准确率最高，为62.79%，即表示共检索回43个单词图像，其中关键词图像包含27个. 而第7、第8两个关键词图像召回率在10个查询关键词中最高，均为65.21%. 即表示共标注23个关键词图像，其中15个目标关键词被召回.

表 2　基于Hu不变矩的关键词检索结果

	1	2	3	4	5	6	7	8	9	10	平均值
标注关键词	40	69	56	48	56	43	23	23	50	45	-
检索关键词	16	27	18	16	24	20	11	15	22	23	-
共检索单词	43	77	46	53	43	40	39	33	55	61	-
准确率/%	37.06	35.06	39.13	30.18	**55.81**	50.00	28.20	45.45	40.00	37.7	39.79
召回率/%	40.02	39.13	32.14	33.33	42.85	46.51	47.82	**65.21**	44.00	51.1	44.21
F值/%	37.33	36.89	35.29	31.67	48.47	48.19	35.47	53.56	41.90	43.3	36.82

表 3 基于改进的Hu不变矩的关键词检索结果

	1	2	3	4	5	6	7	8	9	10	平均值
标注关键词	40	69	56	48	56	43	23	23	50	45	-
检索关键词	20	33	24	22	24	27	15	15	29	28	-
共检索单词	37	64	46	44	43	43	30	31	50	47	-
准确率/%	54.05	51.56	52.17	50.00	55.81	**62.79**	50.00	48.38	58.00	59.5	54.22
召回率/%	50.00	47.82	42.85	45.83	42.85	62.79	**65.21**	**65.21**	58.00	62.2	54.27
F值/%	51.92	49.61	47.05	47.82	48.47	62.79	56.60	55.54	58.00	60.8	53.86

由表3可知，本文改进后的Hu不变矩算法相较于表2中原始的Hu不变矩算法，在准确率和召回率上均有大幅提升. 为进一步验证本文方法的有效性、同时考虑实际扫描中可能存在的问题，通过对数据库图像进行随机旋转、随机缩放、随机遮挡之后的检索效果进行对比，选择以上3种变换的原因是考虑到存在书籍扫描发生倾斜、字体大小不一致和书籍破损与字迹不清等情况，为了模拟以上情况选择了旋转、缩放、遮挡3种变换，其检索效果如图4所示.

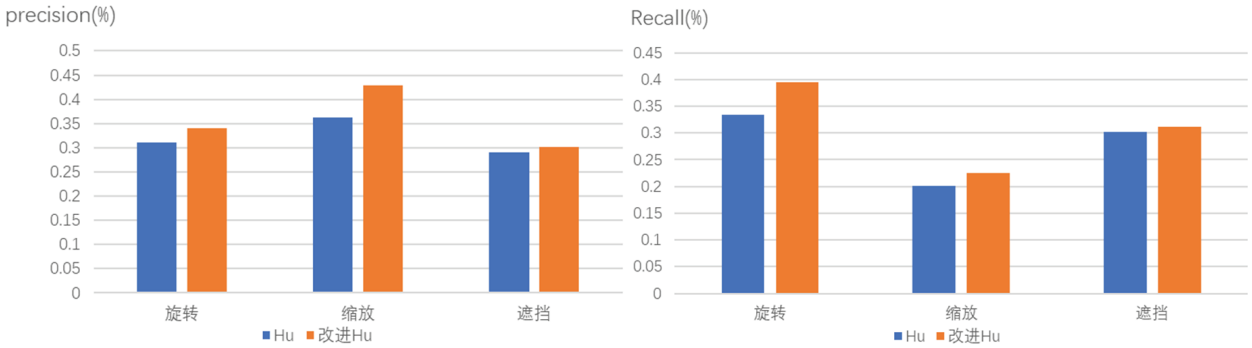


图 4 基于几何变换下的实验结果对比

由图4可知，在经过几何变换后，本文改进后的Hu不变矩算法相对于原始Hu算法在检索准确率和召回率上都有明显提升，从图4左图可以看出经过缩放后的准确率要优于旋转和遮挡，从图4右图可以看出经过旋转后的召回率要优于其它两种变换，而遮挡变换后的检索效果是最差的，这是由于改进后的Hu矩能有效地描述图像的边界和一些不封闭的结构特征信息，可以过滤掉很多相似的图形使得检索准确率更高. 但其整体上效果并不理想，原因在于其特征维度较低未能完全表示出关键词图像的全部图像信息，难免会检索回与目标关键词无关的图像. 因此，可以看出单一特征对图像的表达都具有局限性，无法从不同角度衡量图像，但对不同特征进行融合可以很好地解决这一问题. 因此本文在灰度直方图粗略匹配的基础上，使用改进后的Hu不变矩对粗略匹配检索回的候选单词图像库进行二次精确检索，其效果如表4所示.

表 4 基于灰度直方图+改进的Hu不变矩的关键词检索结果

	1	2	3	4	5	6	7	8	9	10	平均值
标注关键词	40	69	56	48	56	43	23	23	50	45	-
检索关键词	33	69	50	35	28	30	19	20	47	40	-
共检索单词	34	80	83	44	30	44	26	47	52	43	-
准确率/%	**97.05**	86.25	60.24	79.54	93.33	68.18	73.07	42.55	90.38	93.02	78.36
召回率/%	82.50	**100**	89.28	72.91	50.00	69.76	82.60	86.95	94.00	88.88	81.68
F值/%	89.18	92.61	71.93	76.08	65.11	68.96	77.54	57.13	92.15	90.90	78.59

由表4可知，采用二次检索的准确率平均值为78.36%，召回率平均值为81.68%，F值的平均值为78.59%. 其中第1个查询关键词图像准确率在这10个查询关键词中效果最好，为97.05%，即共检索回34个单词图像，其中检索正确的关键词为33个. 而对于召回率而言，第2个关键词检索效果最好，为100%，即标注关键词均被召回.

该阶段是关键词图像的二次精确检索，将灰度直方图特征与Hu不变矩特征进行分层次融合，可以有效地描述图像的边界和空间几何信息，其次能有效地弥补灰度直方图不能描述图像空间位置信息的缺点，通过在粗略匹配形成的候选单词图像集合中进一步实现对关键词的精确检索. 相比表1中灰度直方图粗略匹配效果，其平均准确率提升了31.01%，召回率降低了9.65%. 而相比表3中改进的Hu不变矩的检索结果，其在准确率和召回率上分别提升了24.14%、27.41%. 相比第一阶段的粗略匹配，需要进行特征提取和匹配的单词图像数量已大幅度减少，能有效降低系统复杂度，提高检索效率.

2.3　与现有方法对比

为进一步验证本文提出方法的性能，在相同数据库上和已有的Hu+MB-LBP+OSVM[10]关键词检索方法与模板匹配+HOG+OSVM[11]关键词检索方法做了比较，3种方法的准确率、召回率对比情况见表5. 由表5可知，基于Hu+MB-LBP+OSVM检索方法的平均检索准确率为86.70%，平均召回率为78.30%. 基于模板匹配+HOG+OSVM关键词检索方法的平均检索准确率为91.74%，平均召回率为79.31%. 本文方法在维吾尔文文档图像关键词检索中的平均准确率为78.36%，平均召回率为81.68%. 相较Hu+MB-LBP+OSVM关键词检索方法，本文方法在准确率上降低了8.34%，召回率提升了3.38%. 而相较模板匹配+HOG+OSVM关键词检索方法，本文方法在准确率上降低了13.38%，召回率提升了2.37%. 相较前两种关键词检索方法，本文方法在召回率上有大幅提升，在准确率上还有较大的提升空间.

表 5　基于维吾尔文文档关键词图像检索结果对比

作者	方法	数据集	规模	准确率/%	召回率/%	F值/%
周文杰[10]	Hu+MB-LBP+OSVM	维吾尔文文档图像	115张文档图像10个查询关键词	86.70	78.30	83.80
李静静[11]	模板匹配+HOG+OSVM	维吾尔文文档图像	115张文档图像10个查询关键词	91.74	79.31	84.23
本文方法	灰度直方图+改进Hu矩	维吾尔文文档图像	115张文档图像10个查询关键词	**78.36**	**81.68**	**78.59**

3　结 论

针对维吾尔文单词图像的特点提出了一种特征分层融合的关键词图像二次检索方法，首先利用灰度直方图特征简单快速地过滤掉部分不需要检索的单词图像，其次采用改进的Hu不变矩特征对粗略匹配检索回的单词图像库进行二次精确检索. 通过在粗略检索召回率略有损失的情况下极大提高图像检索的准确率，并通过实验证明该方法的有效性.

随着对关键词检索研究的深入，维吾尔文文档图像的关键词检索依然是一个具有挑战性的研究方向，尤其是关键词图像特征选取的研究，因此我们下一步工作将更加深入分析维吾尔文单词图像的特点并通过寻找新的特征进行融合来提高检索效果. 本文的研究主要是建立在纯文本的印刷体维吾尔文文档图像上，而现实中的文档图像还包含表格、图片等信息，因此，需增加具有复杂布局的文档图像以及手写体的文档图像，同时单词切分方法也有待改进提高，实验中的测试关键词数目以及文档图像规模还会继续扩大，进而降低偶然因素对检索结果的影响.

参考文献：

[1] LEE Y K, SONG J, WON Y. Improving personal information detection using OCR feature recognition rate[J]. The Journal of Super Computing, 2019, 75(4): 1941-1952.

[2] MANMATHA R, HAN C, RISEMAN E M. Word spotting: a new approach to indexing handwriting[C]. Amherst: Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 1996: 631-637.

[3] MANMATHA R, HAN C, RISEMAN E M, et al. Indexing handwriting using word matching[C]. Bethesda: Proceedings of the First ACM International Conference on Digital Libraries, ACM International Conference on Digital Libraries, 1996: 151-159.

[4] ROTHFEDER J L, FENG S, RATH T M. Using corner feature correspondences to rank word images by similarity[C]. Madison: 2003 Conference on Computer Vision and Pattern Recognition Workshop, IEEE, 2003, 3: 30.

[5] VADIVUKARASSI M, PUVIARASAN N, ARUNA P. A frame-work of keyword based image retrieval using proposed Hog_Sift feature extraction method from Twitter dataset[J]. Procedia Computer Science, 2018, 13(4): 1422-1431.

[6] NIAZ H A, AKRAM U, AKBAR U. Word spotting using clustering on extracted DCT and DWT features[C]. Lahore: 2018 International Conference on Engineering and Emerging Technologies(ICEET), IEEE, 2018: 1-4.

[7] 黄祥琳, 高芸, 杨丽芳, 等. 一种基于关键词的中文文档图像检索方法[J]. 中文信息学报, 2017, 4(5): 61-64.
[8] 魏宏喜. 蒙古文古籍图像检索技术研究[D]. 呼和浩特: 内蒙古大学, 2012.
[9] GUO G L, WEI H X, SU X. A case study of BOVW for keyword spotting on historical Mongolian document images[C]. Datong: 2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics(CISP-BMEI), IEEE, 2016: 374-378.
[10] 周文杰. 基于关键词的维吾尔文文档图像检索技术研究[D]. 乌鲁木齐: 新疆大学, 2019.
[11] 李静静, 木特力甫·马木提, 吾尔尼沙·买买提, 等. 基于层级匹配的维吾尔文关键词文档图像检索[J]. 计算机工程与设计, 2020, 41(4): 1062-1069.
[12] 阿丽亚·巴吐尔. 基于局部特征的维吾尔文印刷体复杂文档图像检索研究[D]. 乌鲁木齐: 新疆大学, 2017.
[13] 周文杰, 木特力铺·马木提, 吾尔尼沙·买买提, 等. 基于形态学梯度算法的维吾尔文文档图像单词切分[J]. 计算机工程与设计, 2019, 40(9): 2552-2557.
[14] 欧阳彝华, 黄芳, 周敏. 基于灰度直方图的心脏图像检索[J]. 计算机技术与发展, 2009, 19(9): 125-127+203.
[15] 李顺山, 庄天戈, 陈辉. 基于灰度直方图和互相关方法的医学图像检索[J]. 上海交通大学学报, 2001, 35(5): 694-698.
[16] HU M K. Visual pattern recognition by moment invariants[J]. IRE Transactions on Information Theory, 1962, 8(2): 179-187.

责任编辑：张自强

(上接第 299 页)

[6] TANG N S, QIU S F, TANG M L, et al. Asymptotic confidence interval construction for proportion difference in medical studies with bilateral data[J]. Statistical Methods in Medical Research, 2011, 20(3): 233-259.
[7] PEI Y, TANG M L, GUO J. Test the equality of two proportions for combined unilateral and bilateral data[J]. Communications in Statistics-Simulations and Computation, 2008, 37: 1515-1529.
[8] MA C X, LIU S. Testing equality of proportions for correlated binary data in ophthalmologic studies[J]. Journal of Biopharmaceutical Statistics, 2017, 27(4): 611-619.
[9] TANG M L, TANG N S, ROSNER B. Statistical inference for correlated data in ophthal-mologic studies[J]. Statistics in Medicine, 2006, 25: 2771-2783.
[10] TANG N S, TANG M L, QIU S F. Testing the equality of proportions for correlated otolaryngologic data[J]. Computational Statistics and Data Analysis, 2008, 52: 3719-3729.
[11] TANG N S, QIU S F. Homogeneity test, sample size determination and interval construction of difference of two proportions in stratified bilateral-sample designs[J]. Journal of Statistical Planning and Inference, 2012, 142(5): 1243-1251.
[12] STORER B E, KIM C. Exact properties of some exact test statistics for comparing two binomial proportions[J]. Journal of the American Statistical Association, 1990, 85: 146-155.
[13] BASU D. On the elimination of nuisance parameters[J]. Journal of the American Statistical Association, 1977, 72: 355-366.
[14] LLOYD C J. Exact p-value for discrete models obtained by estimation and maximization[J]. Australian and New Zealand Journal of Statistics, 2008, 50(4): 329-345.
[15] SHAN G. More efficient unconditional tests for exchangeable binary data with equal cluster sizes[J]. Statistics and Probability Letters, 2013, 83(2): 644-649.
[16] TANG M L, TANG N S, ROSNER B. Statistical inference for correlated data in ophthalmologic studies[J]. Statistics in Medicine, 2006, 25: 2771-2783.
[17] SHAN G, MA C X. Exact methods for testing the equality of proportions for binary clustered data from otolaryngologic studies[J]. Statistics in Biopharmaceutical Research, 2014, 6: 115-122.
[18] 陈红. Bootstrap方法在回归分析上的应用[J]. 新疆大学学报(自然科学版), 1999, 16(3): 50-52.
[19] MOU K Y, LI Z M. Homogeneity test of many-to-one risk differences for correlated binary data under optimal algorithms[J]. Complexity, 2021, https://doi.org/10.1155/2021/6685951.
[20] STORER B E, KIM C. Exact properties of some exact test statistics for comparing two binomial proportions[J]. Journal of the American Statistical Association, 1990, 85: 146-155.
[21] BASU D. On the elimination of nuisance parameters[J]. Journal of the American Statistical Association, 1977, 72: 355-366.

责任编辑：赵新科