高技术通讯 2013 年 第23 卷 第6 期:585 ~ 591

doi:10.3772/j.issn.1002-0470.2013.06.006

基于字符基元视觉短语的图像关键字识别①

李 鹏② 崔 刚

(哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)

摘 要 提出了一种利用字符基元视觉短语进行图像关键字识别的方法。该方法通过提取图像关键字的最大稳定极值区域,并进行归一化后得到字符基元。由于通常情况下每个关键字由若干字符基元构成,因此通过采用利用邻接的字符基元构造的视觉短语来提高图像关键字特征描述的可区分性;由于不同的字符基元组合结构可能构成不同的图像关键字,因此基于字符基元相邻关系判断短语几何结构的相似性。此方法不需要对图像进行二值化、布局分析和文本区域定位等预处理操作,具有更好的灵活性和鲁棒性。实验结果表明,此方法对于不同语言的图像关键字识别都具有较高的准确性。

关键词 字符基元,视觉短语,最大稳定极值区域,图像关键字,垃圾邮件图像

0 引 言

图像关键字识别已成为信息检索和过滤领域的关键技术之一[1]。由于不同语言的图像关键字具有不同的轮廓特征,尤其是中文文字具有更复杂的结构,此外,图像关键字通常由多个连通域构成,图像中经常存在噪声干扰,文字呈现无序排列,并且经过尺度和旋转等变换,这些都使得图像关键字的识别和检测面临更大的挑战,从而激发了对图像关键字识别的广泛、深入研究。笔者在借鉴以往研究[2-13]的基础上,已提出了一种基于几何模糊的复杂场景图像关键字识别方法[14]。本研究进行了进一步探索,提出了一种利用字符基元视觉短语进行图像关键字识别的方法。该方法首先提取图像中的最大稳定极值区域(maximally stable extremal region, MSER)[15],并进行归一化,相对于尺度不变特征变换(scale invariant feature transform, SIFT)特征点[16],MSER 区域在图像关键字中的重现率更高,每个归一化后的 MSER 对应一个字符基元。然后利用提取的字符基元构造视觉字典,并将每个基元对应的描述符映射为字典元素索引。通常情况下每个关键字由若干字符基元构成,因此,我们根据 MSER 区域拟合椭圆的邻接特性构造字符基元视觉短语,同一图像关键字中的基元通常位于相同的视觉短语中。最后,由于不同的字符基元组合结构可能构成不同的关键字,因此我们提出一种利用邻接相似性进行弱几何关系一致性判断的方法,从而进一步提高识别的准确性。这种方法不需要对图像进行二值化、布局分析、文本区域定位等预处理操作,因此具有更优的灵活性和鲁棒性。

1 关键字识别

如图 1 所示,基于图像关键字的识别系统流程如下:首先由用户指定检索关键字;然后系统找出对应的图像关键字;最后利用图像关键字特征进行识别,给出那些存在与其匹配的图像。由于目标图像中的关键字可能具有不同的字体、笔画宽度等,所以关键步骤在于图像关键字特征提取与相似性判断。通常,系统无法获取同一关键字的所有不同图像,因此,在有限图像关键字样本的情况下应能够尽可能地检索到最多的相关结果。对于如垃圾邮件图像过滤等应用,光学字符识别(optical character recognition, OCR)无法得到较高的准确率和召回率时,图像关键字过滤是一种更加灵活和有效的方法。本文中,我们将重点关注图像关键字特征提取及匹配方法。

① 国家自然科学基金(61171193)资助项目。

② 男,1984 年生,博士生;研究方向:模式识别,图像处理;联系人,E-mail: lipenggongda@gmail.com

(收稿日期:2012-05-28)

图 1 基于图像关键字的识别系统流程与应用示例

2 字符基元视觉短语

2.1 字符基元提取

如图 1 所示，一个关键字可能对应于多个具有不同字体类型、笔画宽度的图像，但是其彼此之间可能并不存在匹配的尺度不变特征变换（SIFT）等局部特征，所以局部特征点匹配并不直接适用于图像关键字识别。然而，文字的笔画区域具有更稳定的特征。通常，可以根据图像二值化后得到的连通区域提取图像关键字的特征，但这可能丢失一些弱连接，导致匹配准确性下降[8]。本文通过提取最大稳定极值区域（MSER）的方法确定每个图像中的稳定区域。MSER 可以准确地检测图像中的稳定区域，通过 MSER +（较暗的区域）和 MSER-（较亮的区域）可以准确地获取图像中每个稳定区域的内外轮廓。如图 2 所示，黄色表示外轮廓，绿色表示内轮廓，红色表示拟合椭圆（请参阅电子版查看清晰的彩色图像）。然后，根据每个 MSER 的拟合椭圆进行区域归一化，并旋转对齐主方向，即得到本文的字符基元。对于每个字符基元，我们提取 SIFT 描述符作为其特征描述。这样即可提取图像关键字的每个基元，并且使其特征描述具有尺度、旋转和仿射不变性。如图 2 中给出了两个图像关键字样本分解后对应的 MSER 和归一化后的字符基元。由于字符基元“1，一”在归一化后结构特征可区分性下降，因此实际应用中，我们直接将长宽比大于限定阈值的基元作为“1，一”。这样，图像 I 可以表示为字符基元的集合：

$$I = \{m_i \mid m_i = (x_i, y_i, r_i, d_i), 1 \leqslant i \leqslant N\} \tag{1}$$

其中，(x_i, y_i) 为字符基元 m_i 的中心点坐标，r_i 为 m_i 与其所在短语中所有字符基元面积总和之比，d_i 为 m_i 的 SIFT 描述符，N 为 I 中字符基元的个数。需要注意的是，我们只对图像关键字样本中的字符基元计算 r_i，对于待检测的目标图像不计算 r_i，具体原因在下一节中给出。直观上而言，这种方法可以将每幅图像分割为多个 MSER 区域，如果两个图像相似，那么其描述中具有较多的相似字符基元结构。该结构介于 SIFT 特征点和字符的整体形状描述之间，具有更好的灵活性和稳定性。由于较大的 MSER 区域通常不是文本内容，所以我们抛弃目标图像中那些拟合椭圆长轴长度大于图像高或宽 1/3 的区域。

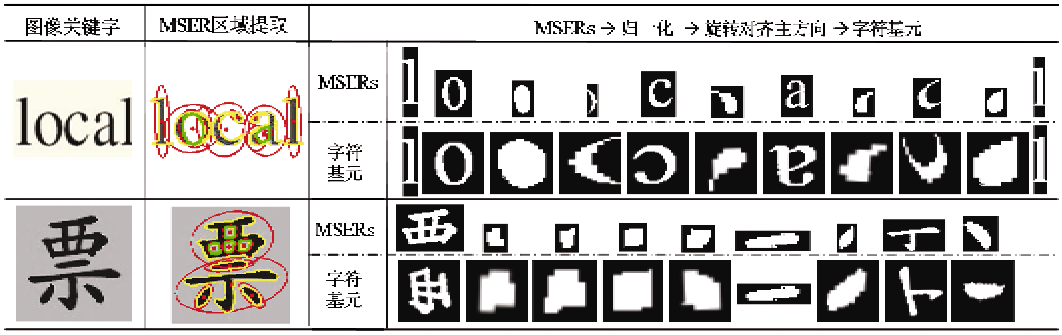


图 2 将图像关键字分解为字符基元集合

2.2 利用字符基元构造视觉短语

对于目标图像，通常可以提取大量的 MSER 区域。在复杂背景下，我们并不知道每个字符基元对应的具体关键字，并且相同的字符基元可能在多个不同的关键字中出现。因此，利用单个字符基元进行关键字检测的可区分性较低，容易导致误匹配。通常，每个图像关键字由多个字符基元构成，如图 2 所示的两个中英图像关键字。观察易知，如果两个基元的拟合椭圆区域相邻，那么这两个基元通常位于相同的关键字中。因此，可以利用不同基元的 MSER 区域拟合椭圆的邻接性构造字符基元视觉短语，从而提高描述可区分性。

首先，定义两个 MSER 区域的邻接性。假设：MSER 区域 m_i 和 m_j 拟合椭圆的长轴和短轴长度分

别为 (l_{m_i}, s_{m_i}) 和 (l_{m_j}, s_{m_j})。如果 $d(m_i, m_j) \leqslant \alpha \cdot (s_{m_i} + s_{m_j})$，则 m_i 和 m_j 直接邻接，记作 $m_i \propto_1 m_j$。其中，$d(m_i, m_j)$ 表示两个区域拟合椭圆的圆心间的距离，α 为调节参数。据此，可以生成所有 MSER 区域之间的直接邻接关系矩阵。邻接具有传递性，如果 $m_i \propto_1 m_j, m_j \propto_1 m_k$，则 m_i 与 m_k 间接邻接，记作 $m_i \propto_2 m_k$。如果 m_i 与 m_j 之间不邻接，记作：$m_i \propto_\times m_j$。最后，我们可以利用迭代的方法确定图像中所有直接或间接邻接的 MSER 区域，从而构成视觉短语集合。这样图像 I 可以表示为

$$I = \{v_i \mid 1 \leqslant i \leqslant K\} \tag{2}$$

其中 v_i 为字符基元视觉短语，并且其中的所有基元彼此直接或者间接邻接。对于同一图像关键字中的各个字符基元，其落入相同短语的可能性较大，反之则较小。经过视觉短语划分，图像中的每个字符基元都将位于唯一的短语中。如图 3 所示，图(a)给出了两个图像关键字样本对应的视觉短语，图(b)给出了从两个实际目标图像中提取的视觉短语。其中红色椭圆为 MSER 的拟合椭圆，同一短语中的直接相邻字符基元通过黄色直线相连。虽然同一关键字可能由多个不同的 MSER 区域提取的字符基元构成，由于彼此临近，因此这些基元都位于同一视觉短语中。但是目标图像中生成的视觉短语中可能存在一些误差。因此，我们并不计算每个 MSER 区域所占的面积比例，而只计算样本图像中的每个区域的面积比，以用于相似性判断。

(a) 图像关键字样本中提取的视觉短语

(b) 实际目标图像中提取的视觉短语

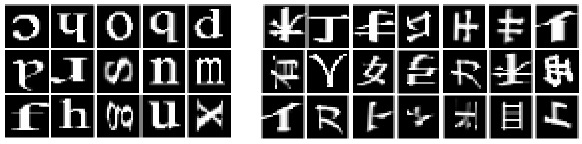
图 3　视觉短语生成

相对于单个字符基元而言，视觉短语由多个邻接的字符基元构成，具有更高的可区分性。同时，有助于降低由于视觉单词量化误差而导致的误匹配，从而提高图像关键字匹配的准确性。接下来的问题是，如何对字符基元短语进行匹配。

2.3　字典构建

直接利用每个短语中所有基元的 SIFT 描述符进行匹配的计算复杂度较高。视觉字典已广泛应用于大规模图像匹配和检索等。因此，我们首先构建字符基元的视觉字典，然后判断每个基元的描述符与视觉单词的距离，最后根据最近距离进行描述符量化。

通常在构建字典元素时，首先需要随机选择大量的样本，然后进行 k-means 聚类，最后以聚类中心作为字典元素。根据文献[17]可知，我们通常无法准确地得知字典规模多大时会得到较低的量化误差和较高的匹配准确性。本文中，由于具有明确的检测目标，并且图像文字通常由一些已知的字符基元构成，为了降低描述符的量化误差，我们通过收集一些已知的具有代表性的字符基元来构造初始字典。对于英文关键字而言，字符基元由 26 个英文字母大小写的不同书写样本得到，而中文字符基元由若干笔画构成。图 4 给出了一些用于构造初始字典的字符基元样例。



(a) 英文字母基元样例　　(b) 中文字符基元样例

图 4　中英文字符基元样例

在描述符量化过程中，通常选择距离最近的字典元素索引作为字符基元的描述。然而，目标图像中还包含大量并不和初始字典元素相似的字符基元。如果仅利用上述字典作为最终字典，那么会存在大量的错误映射，并导致误匹配。所以，在进行字典元素量化时，判断该字符基元到最近的字典元素距离是否小于阈值 τ，如果小于，则量化为该字典元素，否则，将该字符基元作为新的字典元素加入字典中，使得字典具有自学习和补充功能。具体的量化算法如下：

Algorithm 1 Creation of Character Primitive Codebook and descriptor quantization

1: // $W \Leftarrow \{w_1, \cdots, w_m\}$: the list of initial codebook W from representative primitives

2: // $D \Leftarrow \{d_1, \cdots, d_n\}$: the list of descriptor to be quantized

3:　for all descriptors $d_i \in D$ do

4:　　for all primitives of codebook $w_j \in W$ do

```
5:      R_j = d(d_i, w_j)
6:    end for
7:    R_min = min(R_1, R_2, ..., R_m)
8:    if R_min > τ then
9:      W ⇐ W ∪ d_i, d_i = m + 1, m = m + 1
10:   else
11:     d_i = argmin_j(R_1, R_2, ..., R_j, ..., R_m)
12:   end if
13: end for
```

为了提高字典的有效性，我们首先利用图像关键字样本集运行上述算法，以尽可能多地获取有效字符基元。然后，再利用其对目标图像进行字典元素量化。

3 匹配方法

假设 $P = \{p_i \mid 1 \leqslant i \leqslant n\}$ 和 $Q = \{q_i \mid 1 \leqslant i \leqslant m\}$ 为两个短语，其中 $p_i, q_i \in W$，P 为图像关键字样本对应的视觉短语，Q 为目标图像中的一个视觉短语，W 为视觉元素字典。首先，抛弃所有在 Q 中不存在匹配元素的基元；然后，对于剩余的 $p_j \in P$，找到其在 Q 中的对应匹配基元 $q_{j'}$。我们定义 P 与 Q 之间的匹配得分为

$$S(P;Q) = S_m(P;Q) + \lambda \cdot S_g(P;Q) \qquad (3)$$

其中，$S_m(P;Q)$ 为成员相似项，$S_g(P;Q)$ 为几何结构相似项，λ 为加权参数。

成员相似项：通常情况下，如果两个短语中相同的字典元素越多，其相似性越大，否则相似性越小。然而，图像关键字短语中的字符基元具有不同的权重，面积较大的基元具有较大的权重，反之权重则较小。由于实际图像中相同的关键字生成的短语结构可能并不完全相同，因此这里结合目标图像短语 Q 与样本短语 P 匹配的基元的权重判断相似性，即

$$S_m(P;Q) = \sum r_{p_j} \qquad (4)$$

其中 r_{p_j} 为 p_j 的权重，$p_j \in P \cap Q$。

几何结构相似项：相同字符基元的不同几何结构可能构成不同的文字。如图 5 所示，每组关键字图像中都包含一些相同的字符基元，但这些基元之间却具有不同的几何关系结构。英文关键字中的不同字母排列对应不同的单词，笔画结构的不同几何关系也可能对应不同的汉字。因此，必须对两个短语中匹配的字符基元之间的几何关系进行验证。

图 5 相同字符基元的不同几何结构特征

本文中，我们利用不同字符基元之间的位置邻接关系进行非严格性几何结构相似性判断。对于样本图像视觉短语中的任意两个基元，如果其在样本图像中直接相邻，那么目标图像中与其对应的匹配基元也应直接相邻，或者都不直接相邻，即匹配基元间应具有相同的邻接关系。据此，两个短语的几何结构相似项可以表示如下：

$$S_g(P;Q) = \sum \sum e(p_i, p_j, q_{i'}, q_{j'}) \cdot (r_{p_i} + r_{p_j}) \qquad (5)$$

其中：r_{p_i}, r_{p_j} 分别为 p_i 和 p_j 的权重；$p_i, p_j \in P \cap Q$。$e(p_i, p_j, q_{i'}, q_{j'})$ 表示短语 P 中的两个基元与其在 Q 中匹配的两个基元之间的邻接关系是否相同，且

$$e(p_i, p_j, q_{i'}, q_{j'}) = \begin{cases} 1 & \text{if } p_i \propto_1 p_j, q_{i'} \propto_1 q_{j'} \parallel p_i \propto_2 p_j, q_{i'} \propto_2 q_{j'} \\ 0 & \text{其他} \end{cases} \qquad (6)$$

如果两个短语中对应匹配的基元之间具有相同的邻接性，那么则几何关系相似性高，否则低。易知，这种相对位置关系对于图像的旋转、尺度变换都具有不变性。本文方法可以准确地提取字符图像的每个稳定区域，并且特征描述具有尺度、旋转和部分仿射不变性。利用每个短语中不同基元之间的邻接特征进行弱几何关系判断，可以进一步提高匹配的准确性。

为了加速匹配过程，我们利用两个倒排文件对短语进行索引，第一个倒排文件利用字典元素进行索引，保存其在每幅图像的短语中的出现情况。这样对于每次检索，根据索引结构可以很快地找到其与每幅图像中的相似短语，并得到匹配的视觉基元。第二个利用图像标号进行索引，保存每个图像中每个短语内元素的相对位置关系，用于对两个视觉短语之间匹配的元素的几何关系进行判断。

4 实验验证与分析

本节首先给出在测试图像库中的实验结果，然后给出两个中英文关键字的识别样例。

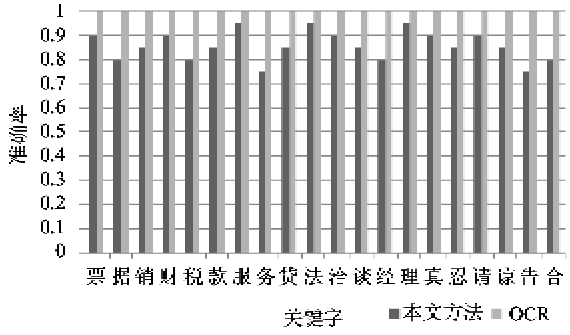
我们利用两个图像库测试本文方法对于不同语言关键字的识别性能:(1)根据垃圾邮件中经常使用的中文关键字生成50幅图像,每幅图像中的关键字具有不同字体类型、笔画宽度,并且经过旋转、尺度等变换;(2)利用相同方法构造50幅包含若干英文单词的图像。

对于每个图像库,我们都选取20个关键字进行测试。对于每个关键字,我们利用对应的三个图像关键字样本进行识别,并与Google的文字识别工具Tesseract-OCR[18]识别性能进行对比。最后,给出不同方法的识别准确率(P)和召回率(R),其中:

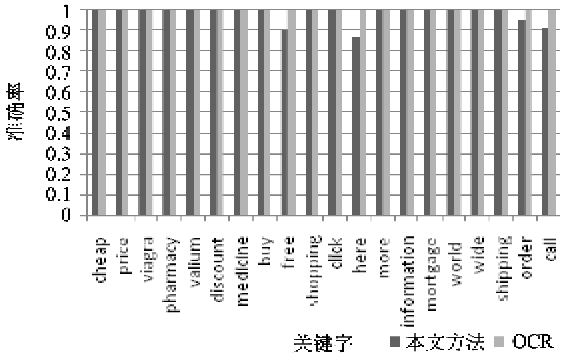
$$P = \frac{\text{正确匹配的目标数量}}{\text{所有判断为匹配的目标数量}} \times 100\% \quad (7)$$

$$R = \frac{\text{正确匹配的目标数量}}{\text{用于测试的所有相关目标数量}} \times 100\% \quad (8)$$

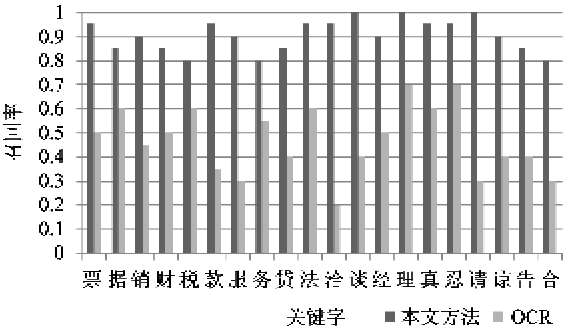
图6(a)和图6(b)给出了本文方法与OCR识别结果的准确率。由结果可知,本文方法和OCR在两个测试集中的识别准确率都较高。OCR识别出



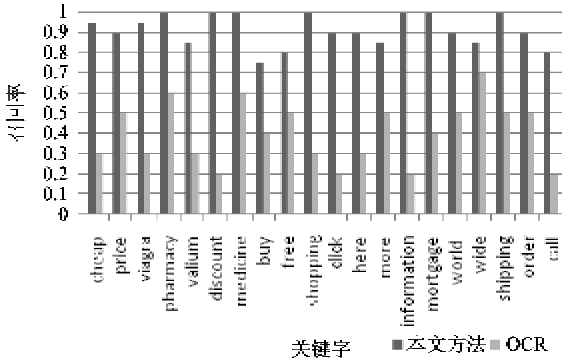
(a) 中文关键字识别准确率对比



(b) 英文关键字识别准确率对比



(c) 中文关键字识别召回率对比



(d) 英文关键字识别召回率对比

图6　本文方法与OCR的识别性能对比

的中英文关键字结果均正确。本文方法对于中文关键字的平均识别准确率约为85%,而英文关键字为98%。由于英文关键字具有更好的结构特征,并且由固定的英文字母构成,而中文关键字可能分割为不同的字符基元。因此,中文关键字的识别准确率要低于英文关键字。由图6(c)和图6(d)可知,与OCR相比,本文方法具有更高的召回率。由于受噪声干扰、排版杂乱等影响,OCR的召回率较低,即大量的关键字没有正确识别。对于英文关键字,OCR通常还难以区分如"1, l(el), i","a, o, 0"等。OCR在两个测试图像库中的召回率约分别为46%和40%,而本文方法的召回率约分别为90%和91%。可知,对于杂乱结构中的图像文字,大量关键字OCR工具无法准确识别,而本文方法则具有较好效果。由于测试图像中的部分关键字与用于匹配的图像关键字样本的字符基元分割结果存在差异,所以导致了部分关键字的漏检。

图7分别给出了本文方法用于真实的中文垃圾邮件图像和部分英文测试图像识别的例子,检索关键字分别为"票"和"viagra"。这里给出了利用一个图像关键字样本识别的结果。每幅图像中利用相同颜色的圆标示同一短语中的各个基元的位置。可知,本文方法可以有效地检测图像中的对应关键字。由于汉字具有复杂的结构,所以图像中一些字体较

图 7 两个图像关键字的识别结果样例

小的关键字并没有检测到。由于这些文字字体较小,在进行 MSER 区域检测时进行了过滤,因此导致了漏检。此外,部分关键字与用于检索的图像关键字样本具有不同的连接结构。这种情况下,对于相同的关键字,需要利用多个图像关键字样本进行识别,以保证较高的识别准确率和召回率。一般情况下,要求同一关键字对应的样本具有不同的区域分割结构。在英文测试图像中,由于存在一些较大的 MSER 区域,导致多个英文单词通过这些较大的区域连接为同一个视觉短语,因此可见在其中两副图像中发现的识别区域较大。

我们在 Matlab R2010b 中实现了本文方法。机器配置为 Intel® Core(TM)2 Duo CPU 2.0GHz, 2GB 内存,操作系统为 Windows Vista。利用倒排文件进行图像特征索引后,对于特定图像关键字“票”的检索时间约为 120ms,具体时间随测试图像中相关短语数量的不同而变化。由于 Matlab 主要用于算法原型验证,对于循环控制类程序执行效率较低。因此,可以通过利用 C/C++ 来进一步提高执行效率。

5 结 论

图像关键字识别作为信息过滤与检索领域的关键技术之一,可以弥补 OCR 应用中的不足。但图像中的文字变化多样,使得图像关键字识别仍然面临很多问题。本文提出了一种利用字符基元视觉短语进行特征描述,并利用元素和几何关系进行相似性判断的方法。实验结果表明,本文方法相对于 OCR 具有更优的召回率。特别是对于类似垃圾邮件图像过滤等应用,OCR 往往不能达到较好的识别效果,本文方法则具有更好的灵活性和更优的召回率。但是,当出现字符基元分割差异较大时,本文方法存在漏检现象。针对这种问题,可以通过收集多样的图像关键字样本用于提高识别的性能。此外,垃圾邮件图像经常借用如 CAPTCHA 技术[13],对图像中的文字内容添加各种干扰,以对抗识别系统,从而降低其过滤的准确性。所以,对于复杂噪声背景下的形变图像文字的识别仍然具有很大挑战性,这些也是我们下一步的工作重点。

参考文献

[1] Murugappan A, Ramachandran B, Dhavachelvan P. A survey of keyword spotting techniques for printed document images. *Artificial Intelligence Review*, 2011, 35(2):119-136

[2] Constantinopoulos C, Meinhardt-Llopis, Liu Y, et al. A robust pipeline for logo detection. In: Proceedings of the IEEE International Conference on Multimedia and Expo, Barcelona, Spain, 2011. 1-6

[3] Li L, Jiang S Q, Huang Q M. Multi-description of local interest point for partial-duplicate image retrieval. In: Proceedings of the International Conference on Image Processing, Hong Kong, China, 2010. 2361-2364

[4] Bai S, Li L, Tan C L. Keyword spotting in document images through word shape coding. In: Proceedings of the 10th International Conference on Document Analysis and Recognition, Barcelona, Spain, 2009. 331-335

[5] Aouadi N, Afef K. Word spotting for Arabic handwritten historical document retrieval using generalized hough transform. In: Proceedings of the 3rd International Conference on Pervasive Patterns and Applications, Rome, Italy, 2011. 67-71

[6] Roy P P, Ramel J, Ragot N. Word retrieval in historical document using character-primitives. In: Proceedings of the International Conference on Document Analysis and Recognition, Beijing, China, 2011. 678-682

[7] Rusinol M, Aldavert D, Toledo R, et al. Browsing heterogeneous document collections by a segmentation-free word spotting method. In: Proceedings of the International Conference on Document Analysis and Recognition, Beijing, China, 2011. 63-67

[8] Leydier Y, Lebourgeois F, Emptoz H. Text search for medieval manuscript images. *Pattern Recognition*, 2007, 40(12):3552-3567

[9] Leydier Y, Ouji A, Lebourgeois F, et al. Towards an omnilingual word retrieval system for ancient manuscripts, *Pattern Recognition*, 2009,42(9):2089-2105

[10] Gatos B, Pratikakis I. Segmentation-free word spotting in historical printed documents. In: Proceedings of the 10th International Conference on Document Analysis and Recognition, Barcelona, Spain, 2009. 271-275

[11] Llados J, Roy P P, Rodriguez J A, et al. Word spotting in archive documents using shape contexts. In: Proceedings of the Iberian Conference on Pattern Recognition and Image Analysis, Girona, Spain, 2007. 290-297

[12] Thayananthan A, Stenger B, Torr P H S, et al. Shape context and chamfer matching in cluttered scenes. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Madison, USA, 2003. 127-133

[13] Mori G, Malik J. Recognizing objects in adversarial clutter: breaking a visual CAPTCHA, In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Madison, USA, 2003. 134-144

[14] 李鹏,崔刚. 基于几何模糊的复杂场景图像关键字识别. 高技术通讯,2013,23(4):379

[15] Matas J, Chum O, Urban M, et al. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 2004, 22(10): 761-767

[16] David G L. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004, 60(2):91-110

[17] Yang J, Jiang Y, Alexander G H, et al. Evaluating bag-of-visual words representations in scene classification. In: Proceeding of the International Workshop on Multimedia Information Retrieval, Bavaria, Germany, 2007. 197-206

[18] Tesseract OCR. http://code. google. com/p/tesseract-ocr/: Google, 2012

Image keyword spotting based on visual phrases of character primitives

Li Peng, Cui Gang

(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001)

Abstract

This paper proposes a new approach for image keyword spotting using visual phrases of character primitives. The approach firstly extracts maximally stable extremal regions from a given image, and then normalizes them into character primitives. Considering that each image keyword is composed of several primitives, it adopts the visual phrases constructed by using the adjacent character primitives to improve the description of image keyword spotting. For different keywords may share the same character primitives but with the different geometric structure, it measures the similarity of geometric structures of phrases based on the adjacent relationship of character primitives. This method does not require the processes of image binarization, layout analysis and text area localization. And it is more flexibly and robust. The experimental results show that this method is with the high accuracy in spotting of different languages' image keywords.

Key words: character primitive, visual phrase, maximally stable extremal region, image keyword, spam image