(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 109086327 B
(45)授权公告日 2022.05.17

(21)申请号 201810716386.4

(22)申请日 2018.07.03

(65)同一申请的已公布的文献号
申请公布号 CN 109086327 A

(43)申请公布日 2018.12.25

(73)专利权人 中国科学院信息工程研究所
地址 100093 北京市海淀区闵庄路甲89号

(72)发明人 柳厅文 李彦增 舒晓波 刘曲
时金桥 李全刚 张水利 亚静

(74)专利代理机构 北京君尚知识产权代理有限公司 11200
专利代理师 邱晓锋

(51)Int.Cl.
G06F 16/958(2019.01)
G06V 30/146(2022.01)
G06T 7/13(2017.01)
G06T 7/155(2017.01)

(56)对比文件
CN 102200966 A,2011.09.28
CN 102200966 A,2011.09.28
CN 108073931 A,2018.05.25
CN 102262618 A,2011.11.30
US 2009148043 A1,2009.06.11
US 2011258195 A1,2011.10.20

审查员 张骞

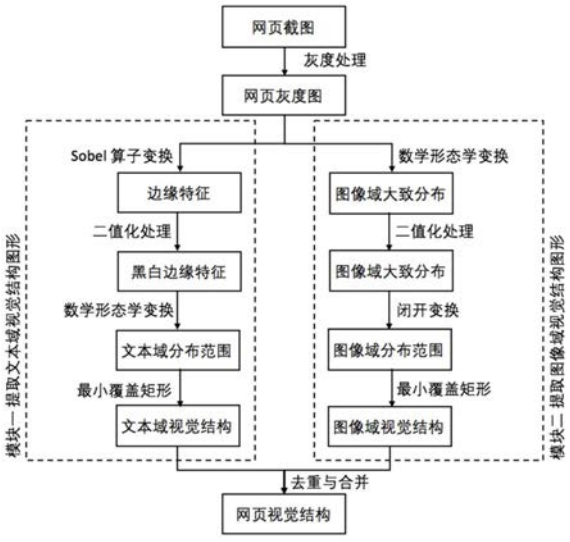
权利要求书2页 说明书4页 附图2页

(54)发明名称

一种快速生成网页视觉结构图形的方法及装置

(57)摘要

本发明涉及一种快速生成网页视觉结构图形的方法及装置。该方法包括:提取网页中的文本域的视觉结构图形;提取网页中的图像域的视觉结构图形;将文本域的视觉结构图形与图像域的视觉结构图形去重与合并,得到网页的视觉结构图形。该装置包括文本域视觉结构图形提取模块、图像域视觉结构图形提取模块、去重与合并模块。本发明抛弃了传统分析方法中网页DOM结构的累赘,仅由网页截图应用图形学方法处理图片,大大降低了算法耗时;本发明采用数学形态学变换,能够快速、准确地分别提取网页中文本域与图像域的视觉结构图形。



CN 109086327 B

1.一种快速生成网页视觉结构图形的方法,其特征在于,包括以下步骤:

提取网页中的文本域的视觉结构图形;

提取网页中的图像域的视觉结构图形;

将文本域的视觉结构图形与图像域的视觉结构图形去重与合并,得到网页的视觉结构图形;

所述提取网页中的图像域的视觉结构图形,包括:

1)对灰度处理后的网页图像直接进行数学形态学变换,使用方形核对灰度图进行多次膨胀;

2)对步骤1)得到的灰度图像进行二值化处理,得到图像域大致分布;

3)对步骤2)得到的黑白图像进行数学形态学处理,进行闭开变换,去除噪点,得到图像域分布范围;

4)对步骤3)得到的图像域分布范围分别计算能覆盖各个图像轮廓的最小矩形,并按照面积进行过滤,将各个矩形区域合并后即可得到最终的图像域视觉特征结构。

2.根据权利要求1所述的方法,其特征在于,所述提取网页中的文本域的视觉结构图形,包括:

1)对灰度处理后的网页图像应用核为1的Sobel算子进行变换,提取出边缘特征;

2)对边缘特征进行二值化处理,将背景色设定为黑色,边缘特征设定为白色;

3)对二值化处理后的边缘特征图像进行数学形态学变换,得到文本域的大致分布范围;

4)对得到的文本域大致分布范围进行边缘提取,得到各个文本域的边缘分布信息;

5)对得到的文本域的边缘分布信息分别计算能覆盖各个边缘的最小矩形,将各个矩形区域合并后即可得到最终的文本域视觉结构图形。

3.根据权利要求2所述的方法,其特征在于,步骤3)所述数学形态学变换包括:

a)对二值化处理后的边缘特征图像进行数学形态学变换,使用方形核对边缘特征进行一次膨胀,使得边缘特征变得更为显著;

b)对膨胀后的边缘特征进行数学形态学变换,使用比上一步中尺寸更大的方形核对膨胀后的边缘特征进行一次腐蚀,消除分割线、图像残余、表格线,得到文本域腐蚀图;

c)对文本域腐蚀图再次应用与步骤a)中相同的数学形态学变换,重复若干次,将腐蚀后的文本域进行强化,直到获得的文本域范围大致不变,由此得到文本域的大致分布范围。

4.根据权利要求1所述的方法,其特征在于,所述将文本域的视觉结构图形与图像域的视觉结构图形去重与合并,包括:通过碰撞算法,找出相互重合的图像域与文本域,将这些图像域去除,仅留下对应的文本域;将剩余的图像域与文本域进行合并,即得到网页的视觉结构图形。

5.一种快速生成网页视觉结构图形的装置,其特征在于,包括:

文本域视觉结构图形提取模块,负责提取网页中的文本域的视觉结构图形;

图像域视觉结构图形提取模块,负责提取网页中的图像域的视觉结构图形;

去重与合并模块,负责将文本域的视觉结构图形与图像域的视觉结构图形去重与合并,得到网页的视觉结构图形;

所述图像域视觉结构图形提取模块采用以下操作提取网页中图像域的视觉结构图形:

1)对灰度处理后的网页图像直接进行数学形态学变换,使用方形核对灰度图进行多次膨胀;

2)对步骤1)得到的灰度图像进行二值化处理,得到图像域大致分布;

3)对步骤2)得到的黑白图像进行数学形态学处理,进行闭开变换,去除噪点,得到图像域分布范围;

4)对步骤3)得到的图像域分布范围分别计算能覆盖各个图像轮廓的最小矩形,并按照面积进行过滤,将各个矩形区域合并后即可得到最终的图像域视觉特征结构。

6.根据权利要求5所述的装置,其特征在于,所述文本域视觉结构图形提取模块采用以下操作提取网页中的文本域的视觉结构图形:

1)对灰度处理后的网页图像应用核为1的Sobel算子进行变换,提取出边缘特征;

2)对边缘特征进行二值化处理,将背景色设定为黑色,边缘特征设定为白色;

3)对二值化处理后的边缘特征图像进行数学形态学变换,得到文本域的大致分布范围;

4)对得到的文本域大致分布范围进行边缘提取,得到各个文本域的边缘分布信息;

5)对得到的文本域的边缘分布信息分别计算能覆盖各个边缘的最小矩形,将各个矩形区域合并后即可得到最终的文本域视觉结构图形。

7.根据权利要求6所述的装置,其特征在于,所述数学形态学变换包括:

a)对二值化处理后的边缘特征图像进行数学形态学变换,使用方形核对边缘特征进行一次膨胀,使得边缘特征变得更为显著;

b)对膨胀后的边缘特征进行数学形态学变换,使用比上一步中尺寸更大的方形核对膨胀后的边缘特征进行一次腐蚀,消除分割线、图像残余、表格线,得到文本域腐蚀图;

c)对文本域腐蚀图再次应用与步骤a)中相同的数学形态学变换,重复若干次,将腐蚀后的文本域进行强化,直到获得的文本域范围大致不变,由此得到文本域的大致分布范围。

8.根据权利要求5所述的装置,其特征在于,所述去重与合并模块通过碰撞算法,找出相互重合的图像域与文本域,将这些图像域去除,仅留下对应的文本域,然后将剩余的图像域与文本域进行合并,即得到网页的视觉结构图形。

一种快速生成网页视觉结构图形的方法及装置

技术领域

[0001] 本发明属于信息技术领域，具体涉及一种快速生成网页视觉结构图形的方法及装置。

背景技术

[0002] 网页视觉结构指的是在排除网页具体内容(包括文字图片)的干扰后，剩下的能够识别网页布局的块状特征，一般由文本域块与图片域块构成。不同的网页通常有着不同的视觉结构，同一个网站中的同一类型网页视觉结构通常类似。

[0003] 由于网页视觉结构排除了具体内容的干扰，可以纯粹地反映出网页元素的分布，因此在信息抽取、恶意网页识别、网页分类中，经常会使用网页视觉结构作为分类器的特征之一。

[0004] 要在大规模的网页分析中应用网页视觉结构特征，首要任务是根据需求提取出特定网页的视觉结构进行分析。传统的视觉结构特征提取方法主要是基于DOM结构的网页视觉块提取方法，通过对DOM树各个节点的tag语义、字体、背景颜色等属性进行分析，得出各个DOM节点所在层级与视觉块，对视觉块进行合并与过滤之后，最终得到网页的视觉结构。最常用的方法是Deng Cai、Shipeng Yu等人提出的VIPS算法。

[0005] 虽然上述方法在生成网页视觉结构的同时，还能得到网页视觉块的层级结构，但由于需要遍历并处理DOM结构，时间复杂度相当高，因此在处理门户网站主页等较大的网页时需要很长的计算时间。但实际上在网页分类等任务中，只需要网页视觉结构图作为特征，不需要具体的视觉块层级结构，因此现有方法的处理方式过于冗余，在处理大规模网页数据时效率太低。

发明内容

[0006] 本发明的目的在于提供一种快速生成网页视觉结构图形的方法及装置，能根据需求快速地提取网页中的视觉结构图形。

[0007] 本发明采用的技术方案如下：

[0008] 一种快速生成网页视觉结构图形的方法，其步骤包括：

[0009] 提取网页中的文本域的视觉结构图形；

[0010] 提取网页中的图像域的视觉结构图形；

[0011] 将文本域的视觉结构图形与图像域的视觉结构图形去重与合并，得到网页的视觉结构图形。

[0012] 进一步地，提取文本域的视觉结构图形的主要步骤包括：

[0013] 1)对灰度处理后的网页图像应用核为1的Sobel算子进行变换，提取出边缘特征；

[0014] 2)对边缘特征进行二值化处理，将背景色设定为黑色，边缘特征设定为白色；

[0015] 3)对二值化处理后的边缘特征图像进行数学形态学变换，使用一定大小的方形核对边缘特征进行一次膨胀，使得边缘特征变得更为显著；

[0016] 4)对膨胀后的边缘特征进行数学形态学变换,使用比上一步中尺寸更大的方形核对膨胀后的边缘特征进行至少一次腐蚀,消除分割线、图像残余、表格线等,由于文本域的边缘特征自身带有的连通性特征,因此不会被完全腐蚀,最终将得到文本域腐蚀图;

[0017] 5)对文本域腐蚀图再次应用与上述第3)步中相同的数学形态学变换,重复若干次,将腐蚀后的文本域进行强化,直到获得的文本域范围大致不变,由此得到文本域的大致分布范围;

[0018] 6)对上述步骤得到的文本域大致分布范围进行边缘提取,得到各个文本域的边缘分布信息;

[0019] 7)对上述步骤得到的文本域的边缘分布信息分别计算能覆盖各个边缘的最小矩形,将各个矩形区域合并后即可得到最终的文本域视觉结构图形。

[0020] 进一步地,提取图像域的视觉结构图形的主要步骤包括:

[0021] 1)对灰度处理后的网页图像直接进行数学形态学变换,使用较小尺寸的方形核对灰度图进行多次膨胀;由于方形核较小,因此占据大部分位置的背景像素会更具优势,将文本消除;而图像由于在图像域范围内为连续分布的像素,因此不会被此处理消除;

[0022] 2)对上述处理得到的灰度图像进行二值化处理,得到图像域大致分布;

[0023] 3)对上述处理得到的黑白图像进行数学形态学处理,进行闭开变换,去除噪点,得到图像域分布范围;

[0024] 4)对上述步骤得到的图像域分布范围分别计算能覆盖各个图像轮廓的最小矩形,并按照面积进行过滤,将各个矩形区域合并后即可得到最终的图像域视觉特征结构。

[0025] 进一步地,将文本域的视觉结构图形与图像域的视觉结构图形去重与合并,包括:第三个模块为合并文本域视觉结构图形与图像域视觉结构图形。通过碰撞算法,找出相互重合的图像域与文本域,这些区域大概率为含有背景的文本。将这些图像域去除,仅留下对应的文本域,最后将剩余的图像域与文本域进行合并,即可得到此页面的视觉结构。

[0026] 与上面方法对应地,本发明还提供一种快速生成网页视觉结构图形的装置,其包括三个模块:

[0027] 第一个模块是文本域视觉结构图形提取模块,负责提取网页中的文本域的视觉结构图形;

[0028] 第二个模块是图像域视觉结构图形提取模块,负责提取网页中的图像域的视觉结构图形;

[0029] 第三个模块是去重与合并模块,负责将文本域视觉结构图形与图像域视觉结构图形去重与合并,最终得到网页的视觉结构图形。

[0030] 在实际应用中,第一个与第二个模块可同时并行,第三个模块需要在第一个与第二个模块都完成之后运行。

[0031] 本发明的技术关键点在于:

[0032] 1)定义了一种快速生成网页视觉结构图形的方法及装置,能根据需求快速地提取该页面的视觉结构图形;

[0033] 2)使用了网页截图图像进行处理,无需考虑网页DOM结构,简化了逻辑与计算成本;

[0034] 3)利用文本边缘特征的连续性与图像的连续性,分别对文本域与图像域应用了不

同的数学形态学变换方法；

[0035] 4)对文本域与图像域的结构范围进行碰撞判断与合并，最终得到网页视觉结构图形。

[0036] 利用本发明提供的方法与设施生成网页视觉结构图形，具有以下优点：

[0037] 1)本发明抛弃了传统分析方法中网页DOM结构的累赘，仅由网页截图应用图形学方法处理图片，大大降低了算法耗时；

[0038] 2)本发明分别定义了两组数学形态学变换的组合，能够快速、准确地分别提取网页中文本域与图像域的视觉结构图形。

附图说明

[0039] 图1是本发明的快速生成网页视觉结构图形的方法的步骤流程图。

[0040] 图2是实施例中快速生成网页视觉结构图形的方法的各步骤的效果图。其中：(a)灰度处理后的网页截图；(b)Sobel算子运算后效果图；(c)数学形态学处理后效果图；(d)文本域最终效果图；(e)对网页截图进行膨胀处理效果图；(f)对图像域图像进行闭开处理效果图；(g)经过碰撞处理后的图像域最终效果图；(h)网页视觉结构效果图。

具体实施方式

[0041] 下面通过具体实施例和附图，对本发明做进一步详细说明。

[0042] 本发明可以应用于任意网页，能根据需求快速地由网页截图提取该页面的视觉结构图形。本实施例以china.com中一条新闻为例，希望得到该网站内容页的视觉结构图形，用于了解该网站的布局结构特征进行后续分析。

[0043] 在该网页中，包含网站导航、右侧导航、底部导航、网页正文文本及图像等区域，结构较为复杂，如使用传统方法对DOM结构进行分析将耗时较长。

[0044] 图1是采用本发明方法对该网站内容页快速生成网页视觉结构图形的步骤流程图。该方法的输入为经过灰度处理的网页截图，输出为网页视觉结构图形。该网页截图如图2(a)所示。

[0045] 1)提取网页文本域视觉结构：

[0046] 第一步：将原网页截图进行灰度处理，并对其应用核为1的Sobel算子变换，提取出图中的边缘特征，如图2(b)所示；

[0047] 第二步：对边缘特征进行二值化处理；

[0048] 第三步：对第二步得到的图像进行数学形态学变换，定义长宽分别为(24,6)的矩形结构核，对上述图像进行1次膨胀变换，让其边缘特征更加明显；

[0049] 第四步：对第三步得到的图像进行数学形态学变换，定义长宽分别为(30,9)的矩形结构核，对上述图像进行2次腐蚀变换，得到文本域腐蚀图；

[0050] 第五步：对第四步得到的腐蚀图像进行数学形态学变换，定义长宽分别为(24,6)的矩形结构核，对上述图像进行3次膨胀变换，让其文本域结构特征更加明显，如图2(c)所示；

[0051] 第六步：对第五步得到的文本域结构图像提取边缘信息，并找出能覆盖各个文本区域的最小面积矩形，将这些矩形合并后即可得到文本域视觉结构图形，如图2(d)所示。

[0052] 2)提取网页图像域视觉结构：

[0053] 第一步：将原网页截图进行灰度处理，并对得到的灰度图进行数学形态学变换，定义长宽分别为(8,8)的矩形结构核，对上述灰度图进行2次膨胀变换，消除图像中的文本，得到只含有图像域的图像，如图2(e)所示；

[0054] 第二步：对第一步得到的图像进行二值化处理；

[0055] 第三步：对第二步中得到的黑白图像进行数学形态学变换，定义长宽分别为(8,8)的矩形结构核，对上述黑白图依次进行一次闭开变换，去除噪点，如图2(f)所示；

[0056] 第四步：对第三步得到的图像域视觉结构图形提取边缘信息，并找出能覆盖各个图像区域的最小面积矩形，并过滤矩形面积小于4000平方像素的矩形区域。将这些矩形区域合并，即可得到图像域视觉结构图形。

[0057] 3)对1)、2)中得到的文本域视觉结构图形与文本域视觉结构图形进行碰撞判断，找出与文本区域重合的图像区域，将这些图像区域删去，如图2(g)所示；

[0058] 4)对3)得到的去重图像域与1)得到的文本域进行合并，得到所需要的网页视觉结构图形，如图2(h)所示。

[0059] 上述实施例表明了本发明的方法是高效、准确的，与传统方法相比，能更加迅速地提取网页视觉结构并生成网页视觉结构图形。

[0060] 本发明另一实施例提供一种快速生成网页视觉结构图形的装置，其包括：

[0061] 文本域视觉结构图形提取模块，负责提取网页中的文本域的视觉结构图形；

[0062] 图像域视觉结构图形提取模块，负责提取网页中图像域的视觉结构图形；

[0063] 去重与合并模块，负责将文本域的视觉结构图形与图像域的视觉结构图形去重与合并，得到网页的视觉结构图形。

[0064] 以上实施例仅用以说明本发明的技术方案而非对其进行限制，本领域的普通技术人员可以对本发明的技术方案进行修改或者等同替换，而不脱离本发明的精神和范围，本发明的保护范围应以权利要求书所述为准。

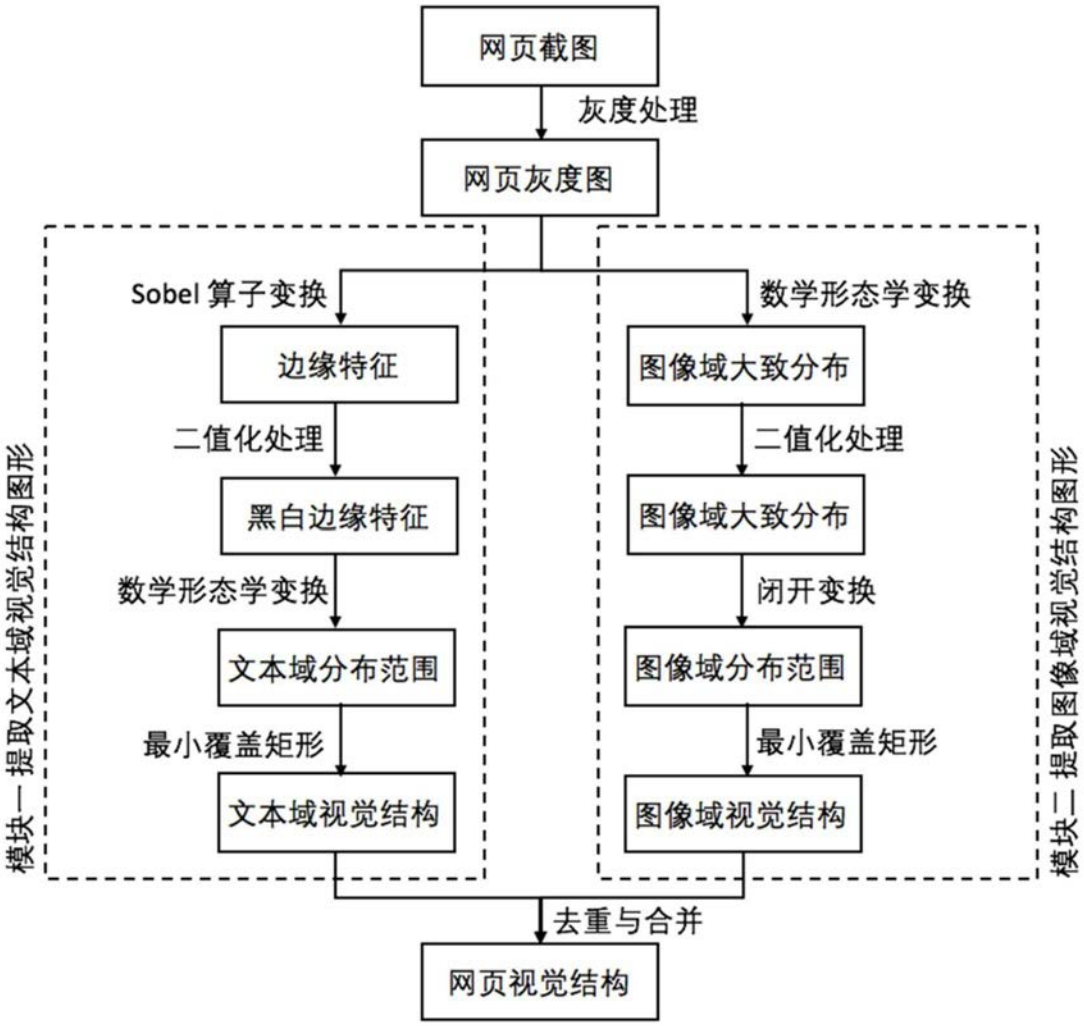

网页截图
灰度处理
网页灰度图
Sobel 算子变换
边缘特征
二值化处理
黑白边缘特征
数学形态学变换
文本域分布范围
最小覆盖矩形
文本域视觉结构
模块一 提取文本域视觉结构图形
数学形态学变换
图像域大致分布
二值化处理
图像域大致分布
闭开变换
图像域分布范围
最小覆盖矩形
图像域视觉结构
模块二 提取图像域视觉结构图形
去重与合并
网页视觉结构

图1

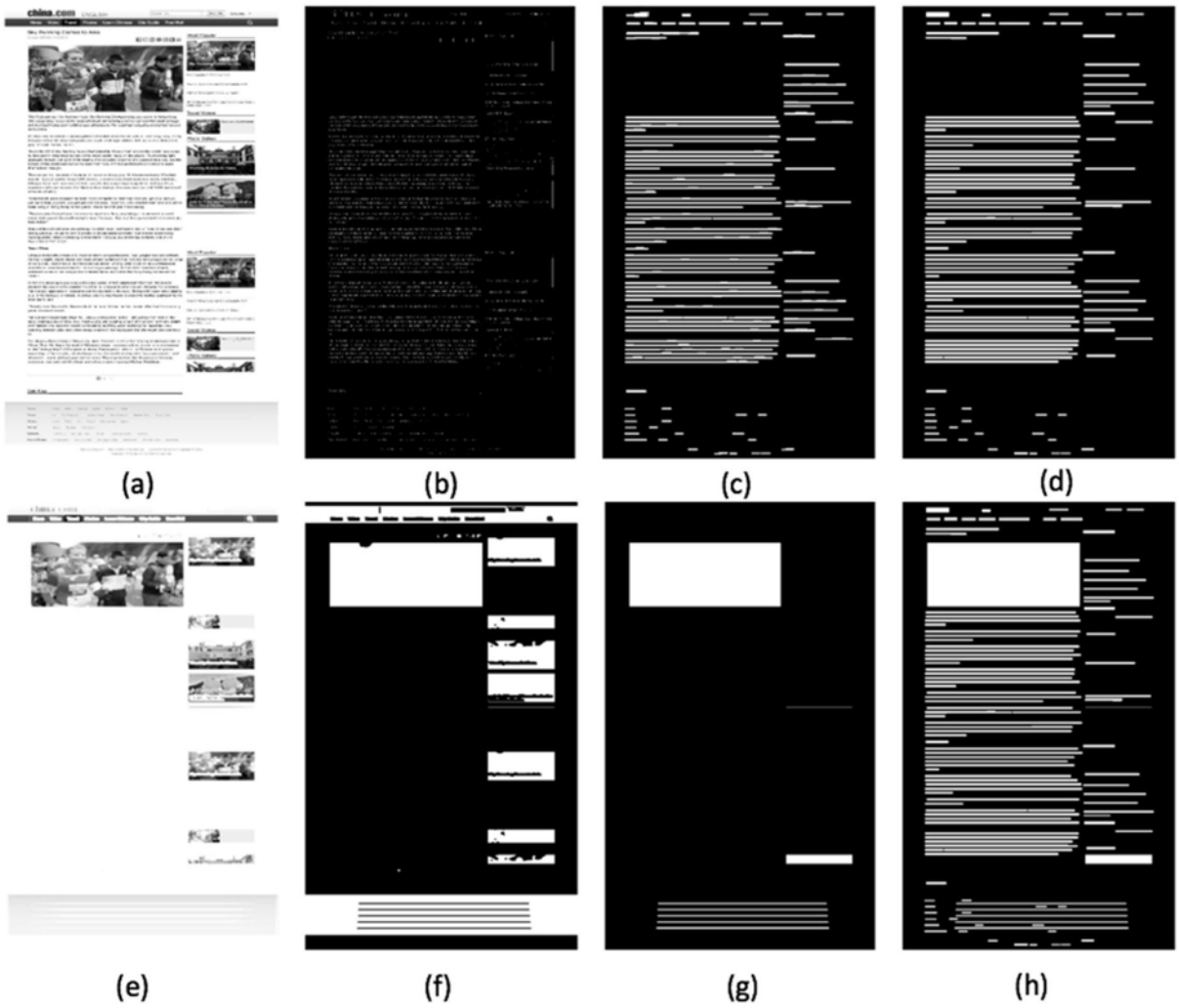


图2