(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 104376317 B
(45)授权公告日 2018.12.14

(21)申请号 201310349738.4

(22)申请日 2013.08.12

(65)同一申请的已公布的文献号
申请公布号 CN 104376317 A

(43)申请公布日 2015.02.25

(73)专利权人 福建福昕软件开发股份有限公司北京分公司
地址 100098 北京市海淀区知春路56号中海实业大厦9层

(72)发明人 周美玲

(74)专利代理机构 北京科龙寰宇知识产权代理有限责任公司 11139
代理人 孙皓晨 朱世定

(51)Int.Cl.
G06K 9/34(2006.01)

(56)对比文件
CN 102456136 A,2012.05.16,maybe.
CN 103186911 A,2013.07.03,不太相关.
CN 102467653 A,2012.05.23,一点相关吧.
CN 103218351 A,2013.07.24,纠偏 去污.
CN 102930267 A,2013.02.13,
US 2004181754 A1,2004.09.16,
杨晓娟等.基于投影法的文档图像分割算法.《成都大学学报(自然科学版)》.2009,第28卷(第2期),
党兴.复杂的中文文档图像版面分析研究.《中国优秀硕士学位论文全文数据库 信息科技辑》.2011,(第1期),

审查员 周循

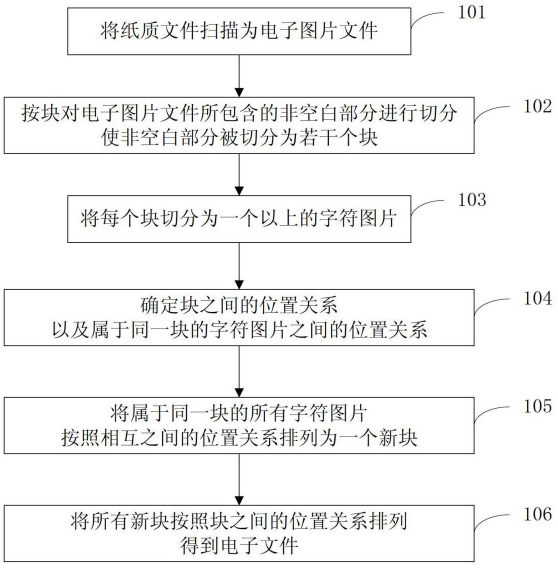
权利要求书1页 说明书5页 附图6页

(54)发明名称

一种将纸质文件转换为电子文件的方法

(57)摘要

本发明涉及一种将纸质文件转换为电子文件的方法。该方法包括:步骤1:用扫描仪将纸质文件扫描为电子图片文件,所述纸质文件为记载在纸张上的文件;步骤2:按块对电子图片文件所包含的非空白部分进行切分,使非空白部分被切分为若干个块;其中,块为行和列中的一种;步骤3:将每个块切分为一个以上的字符图片;步骤4:确定块之间的位置关系以及属于同一块的字符图片之间的位置关系;步骤5:将属于同一块的所有字符图片按照相互之间的位置关系排列为一个新块;步骤6:将所有新块按照块之间的位置关系排列,得到电子文件。本发明能同时提高转换效率以及电子文件与纸质文件内容的相符程度。



CN 104376317 B

1.一种将纸质文件转换为电子文件的方法,其特征在于,该方法包括:

步骤1:用扫描仪将所述纸质文件扫描为电子图片文件,所述纸质文件为记载在纸张上的文件;

步骤2:按块对所述电子图片文件所包含的非空白部分进行切分,使所述非空白部分被切分为若干个所述块;其中,所述块为行和列中的一种;切分原则为:如果所述非空白部分的内容是文字,则所述块为所述文字的每一行或每一列的电子图片;如果所述非空白部分的内容是带边框的表格,则所述块为该表格的电子图片;如果所述非空白部分的内容是不带边框的表格,则所述块为该表格的每一行或每一列的电子图片;如果所述非空白部分的内容是图片,则所述块为该图片的电子图片;

步骤3:将每个所述块切分为一个以上的字符图片;

步骤4:确定所述块之间的位置关系以及属于同一块的所述字符图片之间的位置关系;

步骤5:将属于同一块的所有字符图片按照相互之间的位置关系排列为一个新块;

步骤6:将所有所述新块按照所述块之间的位置关系排列,得到所述电子文件。

2.根据权利要求1所述的方法,其特征在于,在所述步骤1之后,在所述步骤2之前,还包括步骤1-2:旋转所述电子图片文件,使其中的字符处于正向。

3.根据权利要求2所述的方法,其特征在于,在所述步骤1-2中,在旋转所述电子图片文件之前,还包括:删除所述电子图片文件中的污点和划痕。

4.根据权利要求3所述的方法,其特征在于,在所述步骤1-2中,在删除所述电子图片文件中的污点和划痕之前,还包括:放大所述电子图片文件。

5.根据权利要求2所述的方法,其特征在于,在所述步骤1-2中,在旋转所述电子图片文件使其中的字符处于正向之后,还包括:将所述电子图片文件中处于上边距、下边距、左边距及右边距范围内的白边部分切除。

一种将纸质文件转换为电子文件的方法

技术领域

[0001] 本发明涉及将纸质文件转换为电子文件的技术领域，特别是涉及一种将纸质文件转换为电子文件的方法。

背景技术

[0002] 平板电脑、电纸书等技术的出现，使得阅读对象逐渐从纸质文件转换为电子文件，而目前纸质文件浩如烟海，这就需要有将纸质文件转换为电子文件的技术与之相适应来满足读者的阅读需求。

[0003] 常见的将纸质文件转换为电子文件的技术为OCR(Opt ical Character Recognition，光学字符识别)技术，其具体过程为：将纸质文件扫描为电子图片文件；将该电子图片文件切分为多个字符图片，每个字符图片仅包括一个字符；逐个识别每个字符图片中的字符，这其中包括纠错和联想功能以减少错误率；将字符的识别结果按顺序输出，从而得到最终的电子文件。

[0004] OCR技术的核心是对字符图片逐个识别，其判断依据是字符图片的轮廓。由于轮廓相似的字符有很多，因而识别的正确率不高，最终得到的电子文件也就不会太准确。而为了提高识别正确率，OCR技术要花费大量的时间来进行字符识别、查找可疑字符、纠错等处理，因而OCR技术的效率也较低。

发明内容

[0005] 本发明所要解决的技术问题是提供一种将纸质文件转换为电子文件的方法，能同时提高转换效率以及电子文件与纸质文件内容的相符程度。

[0006] 本发明解决上述技术问题的技术方案如下：一种将纸质文件转换为电子文件的方法，该方法包括：

[0007] 步骤1：用扫描仪将所述纸质文件扫描为电子图片文件，所述纸质文件为记载在纸张上的文件；

[0008] 步骤2：按块对所述电子图片文件所包含的非空白部分进行切分，使所述非空白部分被切分为若干个所述块；其中，所述块为行和列中的一种；切分原则为：如果所述非空白部分的内容是文字，则所述块为所述文字的每一行或每一列的电子图片；如果所述非空白部分的内容是带边框的表格，则所述块为该表格的电子图片；如果所述非空白部分的内容是不带边框的表格，则所述块为该表格的每一行或每一列的电子图片；如果所述非空白部分的内容是图片，则所述块为该图片的电子图片；

[0009] 步骤3：将每个所述块切分为一个以上的字符图片；

[0010] 步骤4：确定所述块之间的位置关系以及属于同一块的所述字符图片之间的位置关系；

[0011] 步骤5：将属于同一块的所有字符图片按照相互之间的位置关系排列为一个新块；

[0012] 步骤6：将所有所述新块按照所述块之间的位置关系排列，得到所述电子文件。

[0013] 本发明的有益效果是:本发明中,将纸质文件扫描为电子图片文件,按块对电子图片文件的非空白部分进行切分得到若干个块,然后将块切分为字符图片之后,本发明根据字符图片之间的位置关系将字符图片重新排列为一个新块,根据块之间的位置关系将得到的新块排列为电子文件。因此,本发明无需进行现有的OCR技术中的字符识别、查找可疑字符、纠错、联想等处理,只需利用切分电子图片文件得到的字符图片即可实现转换任务,这大大提高了转换效率,同时,由于本发明利用切分得到的字符图片重新排布得到电子文件,不会引入识别错误,也就大大提高了电子文件与纸质文件内容的相符程度,字符正确率基本可达到100%。

[0014] 在上述技术方案的基础上,本发明还可以做如下改进:

[0015] 进一步,在所述步骤1之后,在所述步骤2之前,还包括步骤1-2:旋转所述电子图片文件,使其中的字符处于正向。

[0016] 进一步,在所述步骤1-2中,在旋转所述电子图片文件之前,还包括:删除所述电子图片文件中的污点和划痕。

[0017] 进一步,在所述步骤1-2中,在删除所述电子图片文件中的污点和划痕之前,还包括:放大所述电子图片文件。

[0018] 进一步,在所述步骤1-2中,在旋转所述电子图片文件使其中的字符处于正向之后,还包括:将所述电子图片文件中处于上边距、下边距、左边距及右边距范围内的白边部分切除。

附图说明

[0019] 图1为本发明提出的将纸质文件转换为电子文件的方法的流程图;

[0020] 图2为本发明扫描得到的一个电子图片文件的示意图;

[0021] 图3为利用本发明对电子图片文件进行旋转后的示意图;

[0022] 图4为利用本发明切除电子图片文件四个边距范围内的白边部分后的示意图;

[0023] 图5为利用本发明按行对电子图片文件所包含的非空白部分进行切分后的示意图;

[0024] 图6为利用本发明将块切分为字符图片后的示意图。

具体实施方式

[0025] 以下结合附图对本发明的原理和特征进行描述,所举实例只用于解释本发明,并非用于限定本发明的范围。

[0026] 本发明提出了一种将纸质文件转换为电子文件的方法,图1为该方法的流程图。如图1所示,该方法包括:

[0027] 步骤101:将纸质文件扫描为电子图片文件。

[0028] 本发明中的纸质文件可以为书籍、画册等任一记载在纸张上的文件。

[0029] 对纸质文件进行扫描从而得到电子图片文件是实现纸质文件电子化的第一步,该步骤可利用扫描仪来完成。

[0030] 步骤102:按块对电子图片文件所包含的非空白部分进行切分,使非空白部分被切分为若干个块。

[0031] 本发明中的块指的是行和列中的一种。

[0032] 电子图片文件由步骤101中的扫描步骤得来，纸质文件中的字符、图、表格等内容必然会在电子图片文件中以某种形式(如以图片的形式等)反映出来，这就对应着电子图片文件中的非空白部分。而除去上述的非空白部分之外，电子图片文件中还必然包含空白部分，例如其上边距、下边距、左边距、右边距范围内的白边部分，等等。

[0033] 该步骤仅对电子图片文件中的非空白部分进行切分，切分结果为若干个块。当然，这里的切分结果也都是电子图片的形式。例如，按照行对非空白部分进行切分，则切分结果为若干个电子图片形式的行。进一步，如果非空白部分的内容是文字，则本步骤得到的切分结果为文字的每一行的电子图片；如果非空白部分的内容是表格，则切分时会区分该表格是带边框的表格还是不带边框的表格，如果是带边框的表格，则将该表格作为一行来处理，即切分结果为该表格的电子图片，如果是不带边框的表格，则将该表格的内容按行来分成块，即切分结果为表格的每一行的电子图片；这里应该注意，本步骤对电子图片文件中内容为图片的部分的切分结果仍为该图片的电子图片，即如果非空白部分的内容为图片，则切分结果仍为该图片的电子图片。按列对非空白部分进行切分的方法与此类此，如果非空白部分的内容是文字，则本步骤得到的切分结果为文字的每一列的电子图片；如果非空白部分的内容是表格，也要区分该表格是带边框的表格还是不带边框的表格，如果是带边框的表格，则将该表格作为一列来处理，即切分结果为该表格的电子图片，如果是不带边框的表格，则将该表格的内容按列来分成块，即切分结果为表格的每一列的电子图片；如果非空白部分的内容为图片，则切分结果仍为该图片的电子图片，这一点与按行进行切分的结果相同。在切分表格时之所以要区分表格是否带有边框，是因为其边框的框线将表格联接为一个整体，不会被分成更小的行或列，因而只能将该表格作为一个整体(即一行或一列)来处理。

[0034] 由于电子图片文件中的空白部分不会与纸质文件中的内容相对应，因而本步骤无需对其进行处理。

[0035] 步骤103：将每个块切分为一个以上的字符图片。

[0036] 步骤102所得到的块只是对电子图片文件中非空白部分的初步切分，事实上，每个块的信息量(即与纸质文件中的内容相对应的内容)仍然较大，所包含的空白部分有时也较多，因而本步骤对每个块进一步进行了切分，得到的结果称为字符图片。由于将块切分成了一个以上的字符图片，因而在多数情况下，每个字符图片所包含的信息量要小于其所属的块，当然，也不排除一个块被切分为一个字符图片，或者块中的所有信息量都被切分到一个字符图片中，其余字符图片全部不包含信息量的情形，在这两种情形中，某个字符图片的信息量与其所属的块相同。

[0037] 本步骤中的字符图片仍是电子图片的形式，其包含的信息不能变化。

[0038] 步骤104：确定块之间的位置关系以及属于同一块的字符图片之间的位置关系。

[0039] 本步骤是确定电子图片文件中非空白部分的布局的步骤。通过确定块之间的位置关系，可确定行与行之间、或者列与列之间的先后顺序，通过确定属于同一块的字符图片之间的位置关系，可以确定同一行的各个字符图片之间的先后顺序。

[0040] 步骤105：将属于同一块的所有字符图片按照相互之间的位置关系排列为一个新块。

[0041] 本步骤是重新排布各字符图片从而得到新块的步骤，排布的规则为步骤104所确定的属于同一块的字符图片之间的位置关系。这样，所得到的新块的内容与相应字符图片所属的块是相同的，而且，由于排布未涉及字符的识别，因而不会出现字符被误读的情况，只要各字符图片的排列顺序正确，各新块中的字符正确率完全可以达到100％。

[0042] 由于每个新块中的各字符图片都来自步骤102所得到的某个块，因而这里的新块与块之间实际上就具有了一一对应关系。

[0043] 步骤106：将所有新块按照块之间的位置关系排列，得到电子文件。

[0044] 本步骤是将步骤105排列得到的新块重新排布的步骤，排布的规则为步骤104所确定的块之间的位置关系。也就是说，本步骤是将新块按照其对应的块在电子图片文件中的顺序来排列，从而得到布局与电子图片文件的布局，同时也是纸质文件的布局一致的电子文件。

[0045] 由此可见，本发明中，将纸质文件扫描为电子图片文件，按块对电子图片文件的非空白部分进行切分得到若干个块，然后将块切分为字符图片之后，本发明根据字符图片之间的位置关系将字符图片重新排列为一个新块，根据块之间的位置关系将得到的新块排列为电子文件。因此，本发明无需进行现有的OCR技术中的字符识别、查找可疑字符、纠错、联想等处理，只需利用切分电子图片文件得到的字符图片即可实现转换任务，这大大提高了转换效率，同时，由于本发明利用切分得到的字符图片重新排布得到电子文件，不会引入识别错误，也就大大提高了电子文件与纸质文件内容的相符程度，字符正确率基本可达到100％。

[0046] 在步骤101之后，在步骤102之前，还可以包括步骤101-102：旋转电子图片文件，使其中的字符处于正向。

[0047] 在步骤101-102中，“字符处于正向”的含义是：如果对字符所处的电子图片文件在屏幕上进行显示，则屏幕上显示的该字符所处的角度与其标准角度完全一致。例如，数字“1”的标准角度为与屏幕或纸面的左右边平行，但在步骤101的扫描步骤中，常常因纸质文件的放置位置不标准而造成扫描得到的电子图片文件发生了一定角度的转动，这样，该电子图片文件中所显示的数字“1”就不再处于其标准角度，而是与电子图片文件(或屏幕)的左右边有了一定的夹角，因而需要在执行步骤102之前对电子图片文件进行旋转，使其中的字符处于正向，以提高步骤102和步骤103切分的正确率。

[0048] 在步骤101-102中，在旋转电子图片文件之前，还可以包括：删除电子图片文件中的污点和划痕。

[0049] 利用该步骤，可以减少或消除污点、划痕等噪音数据对本发明转换正确性的影响，并可以节约转换时间，提高转换效率。

[0050] 进一步，在步骤101-102中，在删除电子图片文件中的污点和划痕之前，还可以包括：放大电子图片文件。

[0051] 放大电子图片文件有利于降低污点、划痕判断的难度，提高判断正确率。

[0052] 此外，在步骤101-102中，在旋转电子图片文件使其中的字符处于正向之后，还可以包括：将电子图片文件中处于上边距、下边距、左边距及右边距范围内的白边部分切除。

[0053] 通过切除电子图片文件中处于上边距、下边距、左边距及右边距范围内的白边部分，可以减少电子图片文件的页面范围，降低后续步骤的工作量，提高转换效率和正确率。

[0054] 图2为本发明扫描得到的一个电子图片文件的示意图，直观看去，图2所显示的内容与扫描前的纸质文件的内容相比，在顺时针方向发生了一定角度的旋转。图中处于上、下、左、右的四条黑线表示该电子图片文件的边界，并无实际意义，图3-图6中各黑线的含义与此相同。

[0055] 图3-图6是对图2电子图片文件进行本发明所述的某些操作步骤后的示意图。其中，图3为利用本发明对电子图片文件进行旋转后的示意图，如图3所示，整个电子图片文件均在逆时针方向相对于图2旋转了一定角度，从而使顶部的图片(标有“Foxit Software”文字及图标、“Company Brochure”文字的黑底图片)及下面的文字都处于各自正向。在图3中，标号301所指示的范围为图3电子图片文件的左边距范围内的白边部分，与此类此，标号302所指示的范围为图3电子图片文件的右边距范围内的白边部分，标号303所指示的范围为图3电子图片文件的上边距范围内的白边部分，标号304所指示的范围为图3电子图片文件的下边距范围内的白边部分。这样，利用本发明切除电子图片文件上边距、下边距、左边距和右边距这四个边距范围内的白边部分后，得到了图4所示的示意图。在此基础上，再按行对电子图片文件所包含的非空白部分进行切分，就得到图5示意图，进而对图5中的各行(包括顶部的图片)进行步骤103所述的进一步切分，就得到图6。由图6可以看出，这里的字符图片可以仅包含一个字符，如将“Company Brochure”切分为15个字母及多个空格，当然，这里的字母和空格仍以电子图片的形式存在。图6中的字符图片还可以包括多个字符，如单词“Solut ion”、“detai ls”等。处于顶部的图片在图6中仍为一个字符图片。

[0056] 由此可见，本发明具有以下优点：

[0057] (1)本发明中，将纸质文件扫描为电子图片文件，按块对电子图片文件的非空白部分进行切分得到若干个块，然后将块切分为字符图片之后，本发明根据字符图片之间的位置关系将字符图片重新排列为一个新块，根据块之间的位置关系将得到的新块排列为电子文件。因此，本发明无需进行现有的OCR技术中的字符识别、查找可疑字符、纠错、联想等处理，只需利用切分电子图片文件得到的字符图片即可实现转换任务，这大大提高了转换效率，同时，由于本发明利用切分得到的字符图片重新排布得到电子文件，不会引入识别错误，也就大大提高了电子文件与纸质文件内容的相符程度，字符正确率基本可达到100％。

[0058] (2)本发明中，在对电子图片文件进行切分之前，还将电子图片文件进行了旋转，使其中的字符处于正向，这有利于提高切分步骤的正确率。

[0059] (3)本发明中，在旋转电子图片文件之前，还删除了电子图片文件中的污点和划痕，可以减少或消除污点、划痕等噪音数据对本发明转换正确性的影响，并可以节约转换时间，提高转换效率。

[0060] (4)本发明通过切除电子图片文件中处于上边距、下边距、左边距及右边距范围内的白边部分，可以减少电子图片文件的页面范围，降低后续步骤的工作量，提高转换效率和正确率。

[0061] 以上所述仅为本发明的较佳实施例，并不用以限制本发明，凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

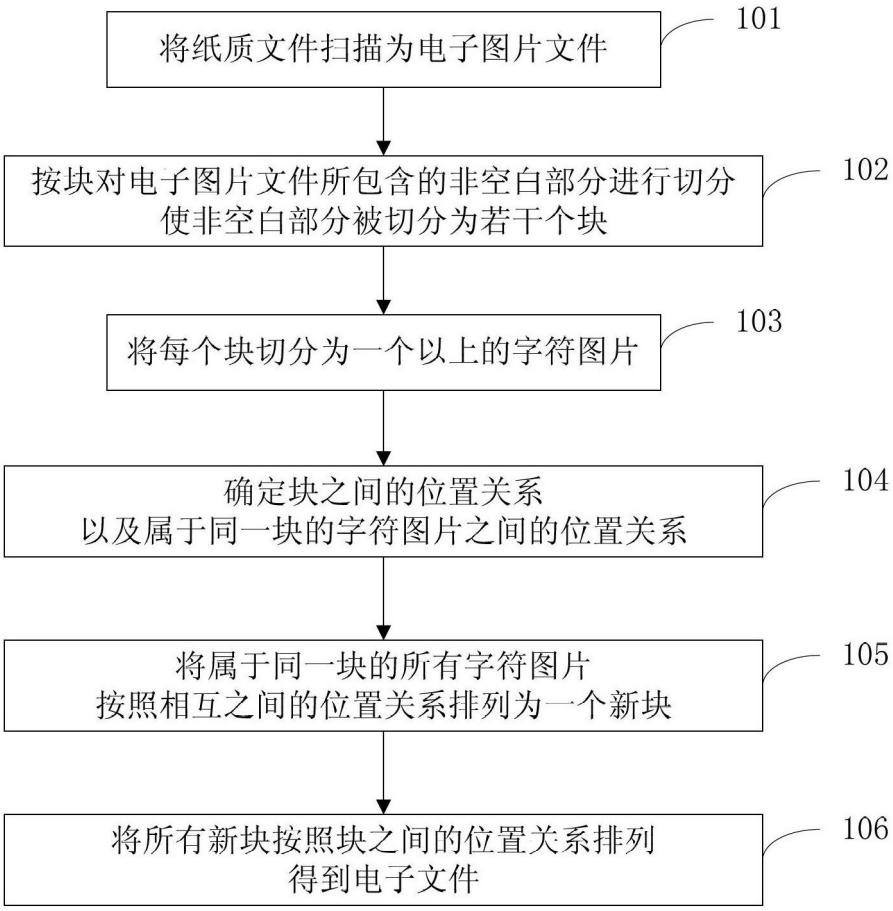

将纸质文件扫描为电子图片文件
101
按块对电子图片文件所包含的非空白部分进行切分
使非空白部分被切分为若干个块
102
将每个块切分为一个以上的字符图片
103
确定块之间的位置关系
以及属于同一块的字符图片之间的位置关系
104
将属于同一块的所有字符图片
按照相互之间的位置关系排列为一个新块
105
将所有新块按照块之间的位置关系排列
得到电子文件
106

图1

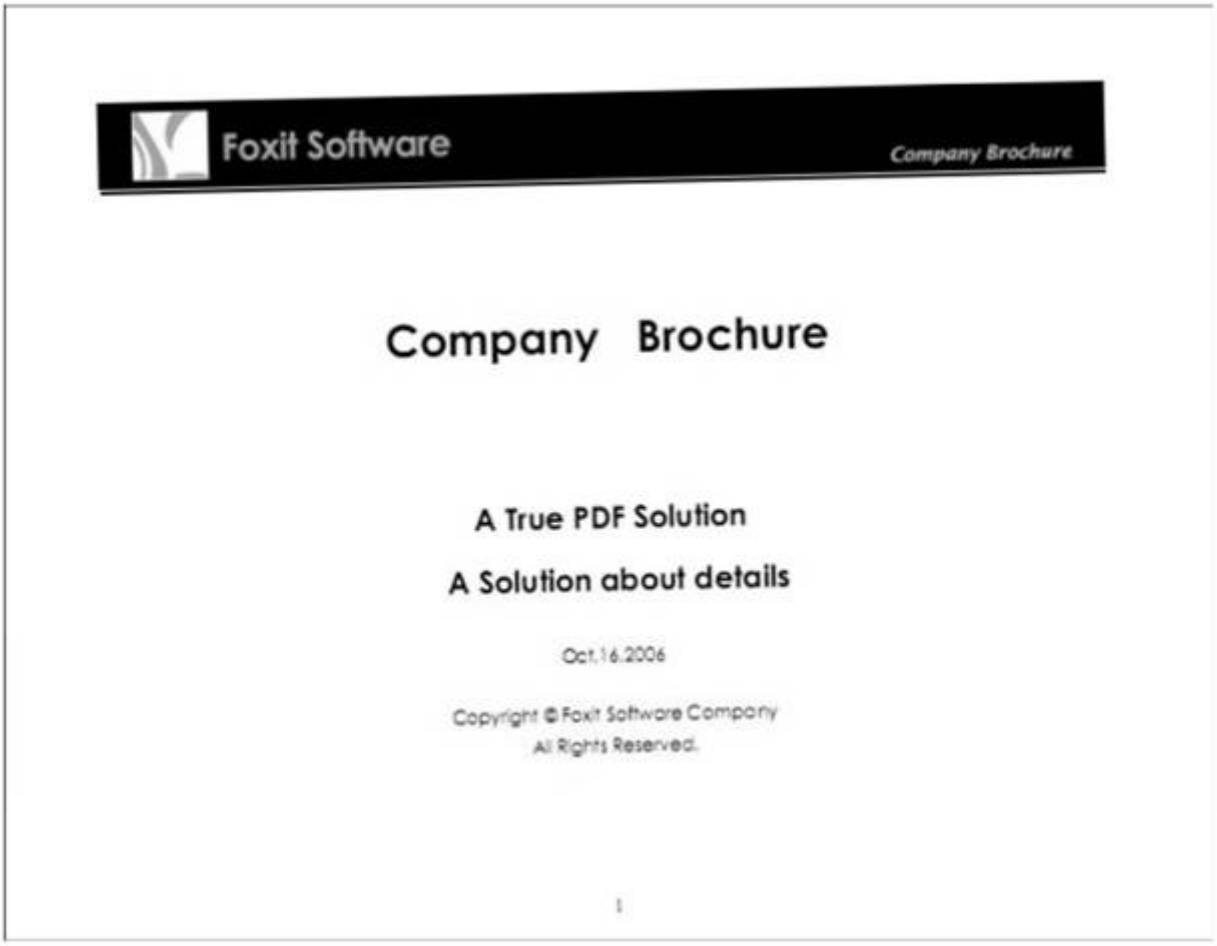

Foxit Software
Company Brochure
Company Brochure
A True PDF Solution
A Solution about details
Oct.16.2006
Copyright © Foxit Software Company
All Rights Reserved.
1

图2


301
302
303
Foxit Software
Company Brochure
Company Brochure
A True PDF Solution
A Solution about details
Oct.16.2004
Copyright © Foxit Software Company
All Rights Reserved.
304

图3


Foxit Software
Company Brochure
Company Brochure
A True PDF Solution
A Solution about details
Oct.16.2006
Copyright © Foxit Software Company
All Rights Reserved.
1

图4


Foxit Software
Company Brochure
Company Brochure
A True PDF Solution
A Solution about details
Oct 16,2008
Copyright © Foxit Software Company
All Rights Reserved

图5


Foxit Software
Company Brochure
Company Brochure
A True PDF Solution
A Solution about details
Oct 16, 2008
Copyright © Foxit Software Company
All Rights Reserved

图6