pdf to jpg without quality loss; gscan2pdf

When I convert a pdf file to bunch of jpg files using

```
convert -quality 100 file.pdf page_%04d.jpg
```

I have appreciable quality loss.

However if I do the following, there is no (noticeable) quality loss:

Start gscan2pdf, choose file-> import (and choose file.pdf). Then go to the temporary directory of gscan2pdf. There are many pnm files (one for every page of the pdf-file). Now I do

```
  for file in *.pnm; do
  convert $file $file.jpg done
```

The resulting jpg-files are (roughly) of the same quality as the original pdf (which is what I want).

Now my question is, if there is a simple command line way to convert the pdf file to a bunch of jpg files without noticeable quality loss? (The solution above is too complicated and time consuming).

/ command-line / image-manipulation / conversion / pdf / imagemagick

asked Apr 22 '11 at 20:02
student
5,460 ● 12 ● 49 ● 100

What is not clear in your questions is whether you talk about text and vector graphics in your pdf, or whether you mean to extract embedded images. – asoundmove Apr 23 '11 at 4:02

4 Answers

It's not clear what you mean by "quality loss". That could mean a lot of different things. Could you post some samples to illustrate? Perhaps cut the same section out of the poor quality and good quality versions (as a PNG to avoid further quality loss).

Perhaps you need to use `-density` to do the conversion at a higher dpi:

```
convert -density 300 file.pdf page_%04d.jpg
```

(You can prepend `-units PixelsPerInch` or `-units PixelsPerCentimeter` if necessary. My copy defaults to ppi.)

Update: As you pointed out, `gscan2pdf` (the way you're using it) is just a wrapper for `pdfimages` (from poppler). `pdfimages` does not do the same thing that `convert` does when given a PDF as input.

`convert` takes the PDF, renders it at some resolution, and uses the resulting bitmap as the source image.

`pdfimages` looks through the PDF for embedded bitmap images and exports each one to a file. It simply ignores any text or vector drawing commands in the PDF.

As a result, if what you have is a PDF that's just a wrapper around a series of bitmaps, `pdfimages` will do a much better job of extracting them, because it gets you the raw data at its original size. You probably also want to use the `-j` option to `pdfimages`, because a PDF can contain raw JPEG data. By default, `pdfimages` converts everything to PNM format, and converting JPEG > PPM > JPEG is a lossy process.

So, try

```
pdfimages -j file.pdf page
```

You may or may not need to follow that with a `convert` to `.jpg` step (depending on what bitmap format the PDF was using).

I tried this command on a PDF that I had made myself from a sequence of JPEG images. The extracted JPEGs were byte-for-byte identical to the source images. You can't get higher quality than that.

edited Apr 22 '11 at 21:57 answered Apr 22 '11 at 20:45

cjm
18.1k ● 5 ● 62 ● 72

+1 I am so glad I didn't submit to the snobbery misreading one of your sentences inspired in me and actually tried pdfimages -- probably the most useful program I have used in months! I'd encourage everyone to try it! – ixtmixilix Apr 22 '12 at 15:45

@ixtmixilix, I'm curious. What did you misread, and how? – cjm Apr 22 '12 at 15:57

Pretty awesome! Solved my day. Thank you! – Geppettvs D'Constanzo Sep 5 '12 at 22:45

`convert` is also impractical for large PDFs. For example, it took 45 GB of memory to process a book of 700 6-megapixel pages. It also took about a thousand times longer than `pdfimages`. – Camille Goudeseune Dec 7 '15 at 22:47

For the other way round, convert images to a pdf, or better, wrap images into a pdf, use img2pdf, here: gitlab.mister-muffin.de/josch/img2pdf (wraps jpg and jpg2000 into a pdf). – erik Mar 15 '16 at 14:10

---

the response from @cjm is correct, but if you like GUI and don't want to render all pdf pages, just to get some image, use gimp.

Open a pdf with gimp an you will get a import window with all pages rendered. Choose whatever pages you want and set resolution to 600 pix/inch (I found 300 too much sharpen in many cases). Save to format you want with "File/export"

Anyway, there must be a flag to select desired pages from command line.

answered Feb 10 '13 at 13:29

albfan
131 ● 5

Looking at the gscan2pdf source code I noticed that it uses pdfimages. So `pdfimages file.pdf page` would result in `page-001.ppm, page-002.ppm` etc.

edited Apr 22 '11 at 20:51 answered Apr 22 '11 at 20:45

student
5,460 ● 12 ● 49 ● 100

What is not clear in your question is whether you talk about text and vector graphics in your pdf, or whether your pdf contains embedded images.

Having read what gscan2pdf is about, my guess is that your pdf files contain (only) embedded graphics.

`convert` essentially "prints" your pdf without regards for what the contents is. Like @cjm suggests, you might want to change the print density. This is the only way to increase quality for vector graphics.

If instead, what you want to do is extract embedded images (much like gscan2pdf seems to do), guessing the density will usually lead to either quality loss or higher quality than required (and waste of disk space). The answer then is to extract the image rather than print the pdf. See this article which basically advocates the use of `pdfimages` in order to extract images without quality loss.

edited Apr 23 '11 at 19:15

answered Apr 23 '11 at 4:07

asoundmove

1,819 ● 1 ● 11 ● 15