Homework 3

Due: 24 September 2018

Problem 3.1

You are evaluating (by hand) a piece of image recognition software that attempts to retrieve images of individual people from an image base and label them as adults or children.

The evaluation corpus has 360 images in it. Of these, 120 are landscapes, 30 are crowded urbanscapes, 60 are logos and graphic designs, 50 are abstract figures and diagrams; 20 mugshots, 50 portraits, and 30 candid photos of individuals. The mugshots are all of adults, but the portraits and candids are each equally divided into pictures of adults and pictures of children. The system yielded the following results:

	Returned image, marked as:	
Actually a:	adult	child
Landscape	2	1
Urbanscape	4	2
Graphic design	1	0
Abstract diagram	0	0
Mugshot	20	0
Adult Portrait	21	1
Child Portrait	3	20
Adult Candid	6	1
Child Candid	0	8

a. First, evaluate the simpler task of retrieving the images of individuals (even if the system marks adult/child incorrectly). What is the system's precision? Recall?

b. How does the evaluation change when you take into account the label the system places on extracted images—that is, images of individuals that are returned but incorrectly labelled are considered incorrect? (This kind of measure is sometimes called "labelled precision" and "labelled recall".)

c. What uses might a system like this be put to?

d. Set aside simple precision and recall for a moment and look at the actual results—are there some subtasks it is better at? How does the composition of the test corpus affect the results? How good is this test corpus

and this evaluation at actually evaluating the system for the intended uses you indicated in part c?

Collaboration policy: group work! If you work with other people on this homework, you can just hand in one copy and put all your names on top. There will be a revision cycle for this.