medr - Medical Surveys Response Predictor

Researcher: Aalok Patwa aalokpatwa21@mittymonarch.com

Purpose

Given a directory of TIFF Scans of Survey Responses, prepare a CSV file medr_predictions.csv with predicted responses.
The predictor is trained for two surveys: UROLOGY ONCOLOGY BENIGN (UOB) AND UROLOGY PROSTATE SYMPTOM (UPS)

Format of medr_predictions.csv

Columns: id,survey, A1,A2,A3,...A13
Rows: One row for each survey in the format below
id is derived from the filename of the TIFF
survey is either 'uob' or 'ups'
A1 is the answer to question 1, in numerical form - ranging from 0 to 6 - the response circled by the respondent.
A2 is similarly the answer to question 2 and thus An is the answer to question n

Method

For each TIFF Scan:
Step 1: Read the scan and use Tesseract to identify the survey: UOB or UPS
Step 2: From the manually measured and coded geometry parameters of the survey:

1) Identify the tables on the page and the question numbers
2) For each question
   2.1) Identify the answerbox locations
   2.2) Crop each answerbox
   2.3) Use trained CNN to predict whether the answerbox was marked or not
   2.4) Formulate the numerical answer after examining all answerboxes.
   Answer is 'NA' if the probability of the prediction is less than the threshold of 0.55

Step 3: Append id, survey, A1, A2, ... A13 in the csv file.

Python3 Libraries

keras

```
tensorflow
pytesseract       # To convert survey scan into text
opencv-python
skimage
PIL
numpy
pandas
csv
os
shutil
random
glob
argparse
matplotlib
sys
time
datetime
```

Package Structure

```
medr/
    medr_predict.py # main program
    utils.py # CNN predictor utilities
    class_list.txt  # list of classes
    ResNet50_model_weights.h5 # trained weights
    medr_predictions.csv # example predictions file
    medr_annotations.csv  # example annotations file
    medr_score.py # to compare medr_annotations against medr_predictions
```

How to Run

1. Install python3 libraries listed above
2. Place all TIF images to run predictions in a directory TIFFDIR (full pathname)
3. Navigate to medr directory and run from there:
   python3 medr_predict.py --tiffdir=TIFFDIR
4. The program will generate “medr_predictions.csv”. If you have medr_annotations.csv - the manually annotated file, you can use medr_score.py to compare and get statistics:
   python3 medr_score.py