Image naming conventions for Trailblazer

For Trailblazer 1 we used a variety of approaches to image naming. Also some times we worked with the original image name as provided by the researcher and at other times the anonymised name as created by us.

The resulting inconsistency made it difficult to manage images and know exactly what image we were referring to.

For future work we need a more organised and standard approach to dataset image naming and management.

Note that the naming convention proposed here is designed to cover pathology images only. We will need a different approach in other areas, eg EM.

Overall approach

For future projects we will use the following approach:

1. The first step when receiving an image dataset from a researcher is to convert it to compressed anonymised jpg files of the required dimensions. We have a Python script that automates this process. It creates a json file containing image metadata that includes the original image name. This will allow us to identify the original image from its new name.
2. The compressed anonymised images are named as described below.
3. We all work with these images and image names (including researchers if possible)
4. Any image variations (eg Illustrator files, annotated images) will be named as described below.

TMA core image naming convention

The images we receive from researchers are named in a variety of different ways. The image names normally encode the slide id and core location. This information needs to be removed in order to anonymise the image.

Examples are:

- GT_CS011_1_1_5_ASMA.jpg
- 7594 Ki67.jpg

We will rename images so that all conform to a common format. This will comprise a number of elements:

1. Tumour type (up to 6 characters)
2. Stain type (up to 6 characters)

3. Dataset identifier (up to 6 characters). Used to distinguish between datasets of same tumour/stain type. This could for instance be researcher id and/or study id.
4. An index number (starting from 1), with leading zeros (number of leading zeros determined by dataset size)
5. File extension (normally jpg)

All name components should be exclusively lower case. Spaces are not allowed in file names – use a dash (minus sign) instead if needed. The different components of the name will be separated by a dash character.

Examples are:

- lung-egfr-gt01-01.jpg
- blad-ki67-ak01-34.jpg
- oesoph-cd8-gt01-0123.jpg (assuming between 1000 and 9999 images in dataset)

These will be the base image names as created by the Python script

Naming of image variations

For some images we will want more than just the base image. We will want annotated and shaded versions, possibly in different sizes. For these we will adopt the following naming convention:

1. Base name (ie tumour-stain-index as in base image)
2. Annotation description if relevant
3. Image size (only if different from original)
4. Different extensions for different file type

Examples

- lung-egfr-gt01-01.ai (Adobe Illustrator version of base image)
- lung-egfr-gt01-01-500.jpg (original image reduced to 500 pixels square)
- lung-egfr-gt01-01-cancer-outline.jpg (original image with added annotations outlining cancer area
- lung-egfr-gt01-01-non-cancer-shaded.jpg (original image with non-cancer area shaded)
- lung-egfr-gt01-01-structure-outline-500.jpg (original image with added structural features outlined, resized to 500 square)

Reference image naming

Reference images (the ones currently displayed across the bottom of the screen in Trailblazer1) will be named in the following way.

1. Tumour type (up to 6 characters)

2. Stain type (up to 6 characters)
3. Dataset identifier (up to 6 characters). Used to distinguish between datasets of same tumour/stain type. This could for instance be researcher id and/or study id.
4. Reference type (ie what is this an image of)
5. An index number (starting from 1), with leading zeros (number of leading zeros determined by dataset size)
6. Image size (only if more than one size is being used)
7. File extension (normally jpg)

Examples:

- lung-egfr-gt01-cancer-01.jpg
- lung-egfr-gt01-non-cancer-08.jpg
- lung-egfr-gt01-stain-bright-05-150.jpg
- lung-egfr-gt01-stain-bright-05-500.jpg

Recording source image for reference and tutorial images

When modifying a reference or tutorial image it can be useful to know the source image that was used to create that image. Incorporating this into the image name will create something rather unwieldy. And from an operational point of view we don't need to know this information – it's only needed when editing an image. This information will therefore be placed in the metadata within the image file (propose origin/source field). This will be invisible in normal use but available when opening an image in Photoshop or Illustrator.

Other tutorial images

There will be other images that we use within the tutorial (eg city analogy, screenshots). These should be named in a way that describes their content or purpose. Again, all lower case with dashes instead of spaces.

Examples:

- mechanic-grid-red-green.jpg
- lung-egfr-city.jpg

Image management

As well as naming the images consistently we should take a standard approach to storing images and the folder structure we use for them.

All images for a given dataset should be placed in a single folder named appropriately to describe the dataset, eg

- lung-egfr
- oesophageal-cd8

Again, all lower case with dashes instead of spaces.in folder names.

Within the main image folder there will be a number of subfolders:

- tutorial (used for all tutorial images, including sample images, annotated versions and practice images)
- reference (for all reference images)

Note that gold standard images will be held within the main folder with all other images. Their status as gold standard will be recorded in an external database and imported into PyBossa