*"Weighted Neighbor Distance using Compound Hierarchy of Algorithms Representing Morphology"*

In CentOS 6.6 Linux box …

**Dependencies to install**

sudo yum install libtiff-devel

sudo yum install fftw-static fftw-devel

sudo yum install fftw-static fftw-devel

sudo yum install libX11-devel libXt-devel libXaw-devel

sudo yum install ImageMagick ImageMagick-devel

**Download zip**

<https://github.com/wnd-charm/wnd-charm>

Download zip directly to Linux box to preserve file permissions

[129 efg localhost 2015-02-04 21:08:43 /home/efg/Software/wnd-charm-master]

./configure

[130 efg localhost 2015-02-04 21:10:03 /home/efg/Software/wnd-charm-master]

make

[131 efg localhost 2015-02-04 21:10:48 /home/efg/Software/wnd-charm-master]

sudo make install

workon Python279

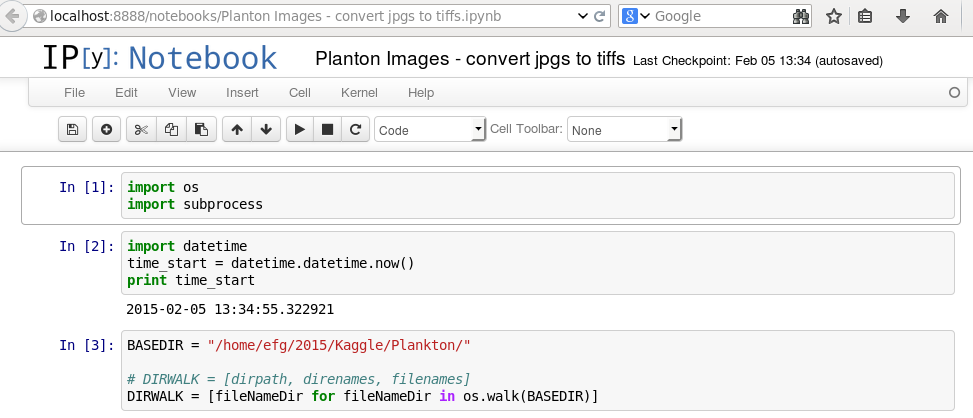
**Convert files: .jpg to .tif**

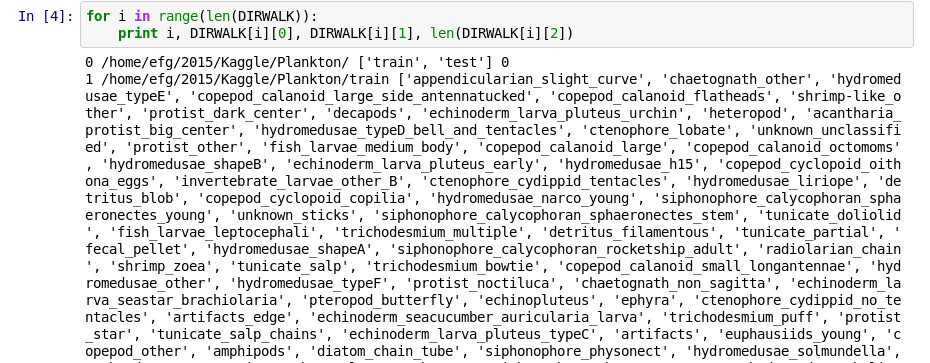
The *wndchrm* program only processes .tif images, so the .jpg files must be converted. *ImageMagick* will be used for this conversion: <http://www.imagemagick.org/>

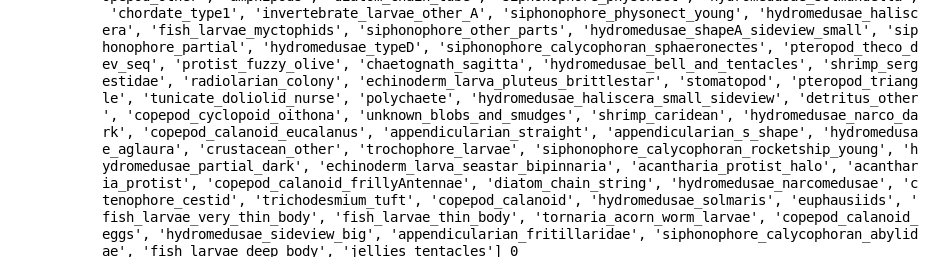
An IPython script was used to document the conversion.

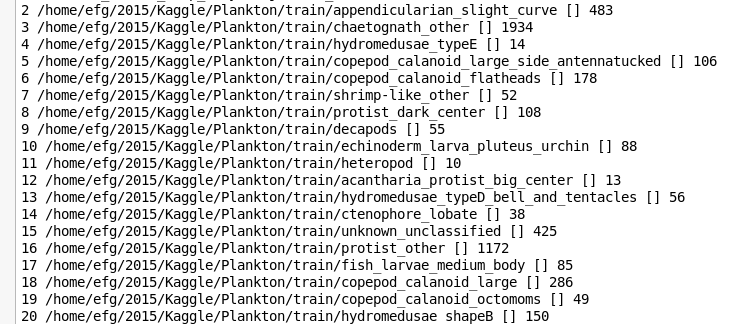
[274 efg localhost 2015-02-05 17:28:21 /home/efg/2015/IPython]

ipython notebook

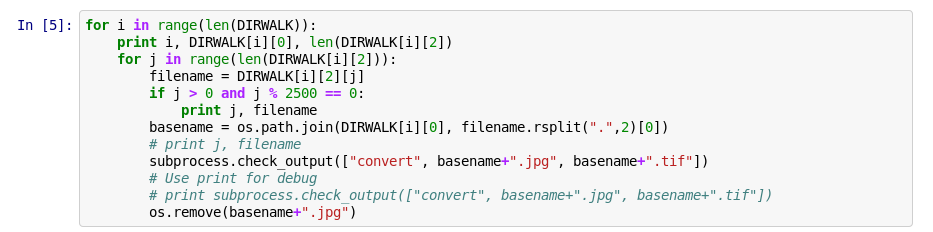








. . .





[295 efg localhost 2015-02-05 19:46:50 /home/efg]

wndchrm

**wndchrm 1.60.** Laboratory of Genetics/NIA/NIH

usage:

======

wndchrm [ train | test | classify | check ] [-mtslcdowfrijnpqvNSBACDTh] [<dataset>|<train set>] [<test set>|<feature file>] [<report\_file>]

<dataset> is a <root directory>, <feature file>, <file of filenames>, <image directory> or <image filename>

<root directory> is a directory of sub-directories containing class images with one class per sub-directory.

The sub-directory names will be used as the class labels. Currently supported file formats: TIFF, PPM.

<feature file> is the file generated by the train command containing all computed image features (should end in .fit).

This filename is a required parameter for storing the output of 'train'

<file of filenames> is a text file listing <image filename>s and corresponding class labels

separated by a <TAB> character (a tab delimited file, or .tsv). Lines beginning with '#' are ignored

<image directory> is a directory of image files. The class cannot be specified so these can only be used as a <test set>.

<image filename> is the full path to an image file. The classes cannot be specified so these can only be used as a <test set>.

<train set> is anything that qualifies as a <dataset>, but must contain at least two (2) defined classes.

An <image filename> or <image directory> cannot define classes.

<test set> is anything that qualifies as a <dataset>. The <train set> will be used to classify the <test set>.

This parameter is required for 'classify' and is optional for 'test' (when doing internal tests of the <train set>

<report\_file> is a report of the test/classify results in html format (must end in .htm or .html).

Image sampling options (require re-computing features):

========================================================

m - Allow running multiple instances of this program concurrently, save (and re-use) pre-calculated .sig files.

This will save and re-use .sig files, making this option useful for single instances/processors as well

R - Add rotations to training images (0,90,180,270 degrees).

t[#][^]C[xR] - split the image into CxC or CxR tiles if R is specified. The default is 1.

If '#' is specified, each tile location is used as a separate dataset (for testing only!).

- If both '#' and '^' are specified only the closest tile is used.

If only C is specified (e.g. -t2), tiling will be CxC (e.g. 2 columns by 2 rows).

- If both C and R are specified (e.g. -t2x3), tiling will be CxR (e.g. 2 columns by 3 rows).

dN - Downsample the images (N percents, where N is 1 to 100)

Sx[:y] - normalize the images such that the mean is set to x and (optinally) the stddev is set to y.

Bx,y,w,h - compute features only from the (x,y,w,h) block of the image.

Image Feature options:

======================

l - Use a large image feature set.

c - Compute color features.

o - force overwriting pre-computed .sig files.

O - if there are pre-computed .sig files accompanying images that have the old-style naming pattern,

skip the check to see that they were calculated with the same wndchrm parameters as the current experiment.

Feature reduction options:

==========================

fN[:M] - maximum number of features out of the dataset (0,1) . The default is 0.15.

v[r|w|+|-][path] - read/write/add/subtract the feature weights from a file.

A - assess the contribution of each group of image features independently.

Classifier options:

===================

w - Classify with wnn instead of wnd.

qN - the number of first closest classes among which the presence of the right class is considered a match.

r[#]N - Fraction of images/samples to be used for training (0,1). The default is 0.75 of

the smallest class. if '#' is specified, force unbalanced training

i[#]N - Set a maximal number of training images (for each class). If the '#' is specified then

the class is ignored if it doesn't have at least N samples.

jN - Set a maximal number of test images (for each class).

nN - Number of repeated random splits. The default is 1.

Nx - set the maximum number of classes (use only the first x classes).

Output options:

===============

s - silent mode. Optionally followed by a verbosity level (higher = more verbose)

p[+][k][#][path] - Report options.

'path' is an optional path to a PHYLIP installation root directory for generating dendrograms.

The optinal '+' creates a 'tsv' directory and exports report data into tsv files.

'k' is an optional digit (1..3) of the specific phylip algorithm to be used.

'#' generates a similarity map of the test images

P[N] - pair-wise distance algorithms for comparing classes

The class probability matrix is the average of marginal probabilities for the images in each class

The similarity matrix is the class probability matrix, where each row is normalized to make the class identity column equal to 1.0

The dis-similarity (i.e. 1.0 - similarity) between two classes can be interpreted as a "morphological distance".

There are two entries in the similarity matrix for each comparison: Class 1 classified as Class 2, and Class 2 classified as Class 1.

N = 1: Use the maximum of the two dis-similarities.

N = 2: Use the average of the two dis-similarities.

N = 3: Use the top triangle only (i.e. Class 1 classified as Class 2)

N = 4: Use the bottom triangle only (i.e. Class 2 classified as Class 1)

N = 5: Use the class probability matrix as a set of coordinates for each class centroid in a "Marginal Probability Space". Use Euclidean distance.

The default method is 5. Method 2 was described in ref [1], and method 5 was described in ref [2].

D[path] - feature file name (.fit file) to save the <dataset> or <train set>.

T[path] - feature file name (.fit file) to save the <test set>.

h - show this note.

Examples:

=========

**train**:

wndchrm train /path/to/dataset/ dataset.fit

wndchrm train -mcl /path/to/dataset/ testset.fit

**test**:

wndchrm test -f0.1 dataset.fit

wndchrm test -f0.1 -r0.9 -n5 dataset.fit testset.fit

wndchrm test -f0.2 -i50 -j20 -n5 -p/path/to/phylip3.65 dataset.fit testset.fit report.html

N.B.: By default, the -r or -i parameters will be used to make a balanced training set (equal number of images per class).

-r#N can be used to override this default, so that the N fraction of each class will be used for training.

If a <test set> is specified, it will be used as the test set for each 'split', but training images will

still be randomly chosen from <train set>)

**classify**:

wndchrm classify dataset.fit /path/to/image.tiff

wndchrm classify -f0.2 -cl /path/to/root/dir /path/to/image/directory/

wndchrm classify -f0.2 -cl -Ttestset.fit dataset.fit /path/to/image/file\_of\_filenames.tsv

N.B.: classify will use -r or -i to train with fewer than all of the images in <dataset>

Unlike 'test', 'classify' will chose the training images in order rather than randomly.

classify will ignore the -n parameter because the result will be the same for each run or split.

The default -r for 'classify' is 1.0 rather than the 0.75 used in 'test'.

**check:** Report which features will need to be computed

wndchrm check path/to/image.tiff

Additional help:

================

A detailed description can be found in: Shamir, L., Orlov, N., Eckley, D.M., Macura, T., Johnston, J., Goldberg, I.

[1] "Wndchrm - an open source utility for biological image analysis", BMC Source Code for Biology and Medicine, 3:13, 2008. <http://ome.grc.nia.nih.gov/wnd-charm/BMC-wndchrm-utility.pdf>

An application of pattern recognition for a quantitative biological assay based on morphology can be found in:

[2] Johnston, J., Iser W. B., Chow, D. K., Goldberg, I. G., Wolkow, C. A. "Quantitative Image Analysis Reveals

Distinct Structural Transitions during Aging in Caenorhabditis elegans Tissues", PLoS ONE, 3:7:e2821, 2008.

If you have questions or problems with this software, please visit our Google code page <http://code.google.com/p/wnd-charm/>

**Using the Command Line Utility**

This paper describes the command-line parameters:  
<http://ome.grc.nia.nih.gov/wnd-charm/BMC-wndchrm-utility.pdf>, p. 5/13

Train

In order to train an image classifier, the first required task

is computing image content descriptors for all images in

the dataset. These numeric values describe the image content

in a fashion that can later be processed by pattern recognition

methods. This step is performed by using a

simple command line described below:

wndchrm **train** [options] *images feature\_file*

where ***feature\_file***is the resulting output file of the image

feature values, ***images***is a path to the top folder where the

images of the dataset are stored, and [*options*] are optional

switches that can be specified by the user. The top folder

should consist of several sub-folders such that each subfolder

contains images of a different class.

The single output file ***feature\_file***contains features for all

classes in the dataset, so that there are no separate files for

the different classes. Therefore, if a new class is added to

the dataset, a new file needs to be created using the same

command line. To avoid re-computing classes that have

already been computed, the user is advised to use the "-m"

switch that will be described later in this section.

Test

Once all image content descriptors are computed, the

dataset can be tested for classification accuracy. This can

be done by using the following command line:

wndchrm test [options] *feature\_file* [report\_file]

where ***feature\_file***is the output file of the train task, and

***report\_file***is an optional html file providing detailed information

regarding the performance of the classifier. This

instruction automatically splits the images of each class

into training and test images, and the effectiveness of the

classifier is determined by the percentage of test images

that are classified correctly using the training images. The

test images are classified by computing the Fisher scores

and assigning the image features with weights. The output

of this command is a confusion matrix, a similarity

matrix, and the accuracy of the classifier (the percentage of

test images that were classified correctly).

When testing an image classifier, the user can determine

the number of images that are used for testing and the

number of images used for training. By default, 75% of

the images of each class are used for the training, and the

remaining 25% are used for testing. The user can change

this ratio by using the "-r" option. For example, "-r0.4"

allocates 40% of the images for testing, and 60% for training.

The allocation of the images to training and test sets

is performed in a random fashion. Users can repeat the

test with several different random splits in a single command

by specifying the "-n" option, followed by the

requested number of splits.

*wndchrm* also allows the user to set the number of training

images per class. This can be done using the "-i" option,

followed by the requested number of training images per

class. The remaining images are used for testing, unless the

"-j" option is used in a similar fashion to set the number

of test images per class. It should be noted that "-i" and "-

j" options override the "-r" value. If only one of these

options is specified, and "-r" is also used, the number of

training or test images per class (the one that is not determined

by "-i" or "-j") will be determined by the "-r" value.

**We suggest using a fixed number of training images per**

**class ("-i") when generating similarity matrices.**

Users can also use different feature files (generated by

using *wndchrm*'s "train" command) for testing and training,

so that instead of splitting a single dataset into training

and test images, one dataset is used entirely for

training while a second dataset is used for testing. This can

be done by **simply specifying two full paths to image feature**

**files. If two files are specified, the first will be used for**

**training and the second for testing.**

Classify

After a classifier is trained and tested, an image can be classified

using the command line:

wndchrm classify *feature\_file image*

where ***image*** can be a full path to the image being classified,

or a folder that contains multiple *images*. If image

points to a specific image file, the output of this instruction

is the predicted class in which the image belongs, as

well as a vector of similarity values to each of the classes

in the dataset. If *image* is a path to a folder, *wndchrm* classifies

and prints the predicted class and similarity vector

for each image in that folder, followed by a brief summary

that specifies the number of images that were classified to

each class and the average similarity vector.

***Changing the Number of Image Features***

Since *wndchrm* is a multi-purpose tool designed to handle

many different image datasets, it uses very many different

image features. However, for a given dataset, not all image

features are assumed to be equally informative, and some

of these features are expected to represent noise. By

default, only the 0.15 images features with the highest

Fisher scores are used. Users who wish to change this setting

can specify the "-f" option in the command line, followed

by the requested portion of the image features to be

used. Changing this value can affect the performance of

the image analysis since in some datasets more image features

may be informative, so that using more features can

contribute to the discrimination between the classes. On

the other hand, in other datasets only few of the image

features provide discriminative information, and using

the non-informative features can add confusion and

degrade the efficacy of the analysis. Since image features

are weighed by their informativeness, the effect of noisy

features is expected to be lower than the effect of more

informative features. However, if very many non-informative

image features are used, their large number can be

weighed against their low Fisher scores, leading to an

undesirable degradation of the performance. **Therefore,**

**the threshold for non-informative features needs to be**

**determined operationally for each type of data, and the**

**0.15 threshold is only a starting point.**

***Image Tiling***

In some cases it may be useful to divide large images of tissues

or cells into several equal-sized tiles. For example, it

has been demonstrated that when each image captures

very many cells, dividing the image into tiles can in some

cases provide better analysis than applying a first step of

global-thresholding cell segmentation [22]. Another

advantage is that using more tiles can improve the effectiveness

of the Fisher scores assigned to the image features,

which are expected to improve as the size of the dataset

gets larger. Using *wndchrm*, this can be done by specifying

the "-t" option followed by the square root of the desired

number of tiles. For instance, "-t3" divides each image

into 3 × 3 tiles.

If segmentation of the subjects (e.g., cell segmentation,

bone segmentation, etc) is required, the user has to apply

a first step of segmentation using a designated utility. The

output of the utility (the segmented subjects) can be used

as input for *wndchrm*, rather than the original images

**Phylip Installation**

<http://evolution.genetics.washington.edu/phylip.html>

make install

**[302 efg localhost 2015-02-05 18:59:30 /home/efg/Software/Phylip/phylip-3.696]**

ll

total 20

drwxr-xr-x. 3 efg efg 4096 Feb 5 18:45 doc

drwxr-xr-x. 2 efg efg 4096 Feb 5 18:48 exe

-rw-r--r--. 1 efg efg 7593 Sep 21 12:19 phylip.html

drwxr-xr-x. 6 efg efg 4096 Feb 5 18:48 src

**Train**

**[298 efg localhost 2015-02-05 19:02:46 /home/efg/2015/Kaggle/Plankton]**

ll

total 7640

drwxrwxrwx. 2 efg efg 7819264 Feb 5 14:33 test

drwxrwxrwx. 123 efg efg 4096 Dec 30 12:19 train

To use four processors:

cd /home/efg/2015/Kaggle/Plankton

wndchrm train -ml /home/efg/2015/Kaggle/Plankton/train Plankton-01.fit

wndchrm train -ml /home/efg/2015/Kaggle/Plankton/train Plankton-01.fit

wndchrm train -ml /home/efg/2015/Kaggle/Plankton/train Plankton-01.fit

wndchrm train -ml /home/efg/2015/Kaggle/Plankton/train Plankton-01.fit

-m: Allow running multiple instances of this program concurrently.  
-l: Use large image feature set.