Automatic Particle Picking Benchmark

User Guide

**Aim and scope**

Picking thousands of particles for performing a 3D reconstruction can be a cumbersome task. Automatic particle picking algorithms aim at alleviating this problem by helping the user to quickly identify those regions in the micrographs corresponding to the kind of particles she is interested in. In this task 3DEM Benchmark Web Site provides datasets helping developers to assess the accuracy of their algorithms with respect to particles manually selected.

Each dataset is composed by three different subsets:

1. Training: a medium number of micrographs is provided along with the coordinates of the manually selected particles. These particles may be used to train the algorithm.
2. Test: a small number of micrographs is provided along with the coordinates of the manually selected particles. This set can be used by developers to validate the interchange protocol with the 3DEM Benchmark Web Site.
3. Challenge: a large number of micrographs is provided and the coordinates of the particles are not provided.

The accuracy assessment process involves downloading the datasets, picking the particles and uploading the coordinates of the selected particles to the 3DEM Benchmark Web Site. Then, the Web Site will report the results using the figures of merit described in the next section. A standard workflow starts training the algorithm (if needed) with the training set, then the test set will be used to validate the interchange format (described in “Data Formats” section) and finally the challenge dataset will be used to assess the accuracy of the algorithm.

In adition to global figures of merit the 3DEM Benchmark Web Site allows to check the results visually (see Figure 1).

Figure 1: Micrograph displaying true solutions (white dots) and the result algorithm selection (green dots).

**Figures of merit**

Let us consider the following matrix of possible outcomes

|  |  |  |
| --- | --- | --- |
|  | **Not selected by the algorithm** | **Selected by the algorithm** |
| **Not present in the oficial solution** | N00 | N01 |
| **Present in the oficial solution** | N10 | N11 |

N11 is the number of true positives (particles that are present in the official solution and were selected by the algorithm). N01 is the number of false positives (particles that were selected by the algorithm but do not correspond to true particles). N10 is the total number of false negatives (particles that were not identified by the algorithm and are present in the official solution).

Then we define the following figures of merit

* Precision rate: this is the percentage estimates the probability a classifier's positive prediction is actually positive. The Precision is defined as Precision=N11/(N10+N11).
* Recall rate: this is the percentage of selected particles that correspond to true particles. The Recall is defined as Recall=N11/(N01+N11)
* False Discovery (FDR): this is the percentage of true particles that is missed by the algorithm. It is defined as FDR=N10/(N10+N11)
* F-measure: this is the harmonic mean of the precision and the recall. It is defined as F-Measure=2 \* Precision \* Recall/(Precision + Recall)

For the true positives we also measure the distance between the manual and algorithmically selected particles. For these distances we compute the average and its histogram resulting in two more figures of merit:

* Distance average
* Histogram

**Uploading procedure**

Once you have the upload file ready, you must login to the web page. If you do not have a username and password, select the register button and fill the form with your personal information. A mail will be sent to the mail you have introduced. To finish the registration process click on the link that you will find in the mail.

Next, go to the benchmarking task web page you want to submit and click on the upload associated to the dataset you have used in your experiments. Introduce a title and, optionally, a short description telling which parameters and the kind of algorithms you have executed to obtain results. Attach the file and, finally, select the mode you want to submit the results. You can choose among:

1. Test: This upload will only be accessible to you when you visit the web site as a registered user. Later, from ‘My Benchmark’ web page, you will be able to delete the upload, edit the comments and title as well as modifying the status to one of the next two.
2. Public-Anonymous: Your results can be requested by any user, however he will not know which user has submitted them. From ‘My Benchmark’ section, you, only, will be able to edit the comments and title and to change the status to the third short.
3. Public-[Username]: You upload your results as an identified user and they will be accessible to any user. From ‘My Benchmark’ section, you will be able to edit the comments and title.

After clicking on the ‘Go’ button, you will see some figures of merit comparing your results with the gold standard/s which would have been recorded on the database. If, finally, you do not want to submit your upload you still can cancel the upload process. Otherwise, press the ‘Ok’ button; the information will be inserted in the system and a confirmation message with the upload id will be shown.

**Data formats**

Images are written in MRC image format.

The uploaded coordinates and the true coordinates (Gold Standard) must be written in the following XML format:

**<?xml version="1.0" encoding="ISO-8859-1"?>**

**<!DOCTYPE particlepicking PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN" "http://i2pc.cnb.csic.es/3dembenchmark/dtd/particlePicking.dtd">**

**<particlepicking>**

**<micrograph id="0001">**

**<coordinate x="*X*" y="*Y*"/>**

**<coordinate x="*X2*" y="*Y2*"/>**

**….**

**<coordinate x="*Xn*" y="*Yn*"/>**

**</micrograph>**

**<micrograph id="0002">**

**<coordinate x="*X’*" y="*Y’*"/>**

**<coordinate x="*X’2*" y="*Y’2*"/>**

**….**

**<coordinate x="*X’m*" y="*Y’m*"/>**

**</micrograph>**

**….**

**<micrograph id="RRRR">**

**<coordinate x="*X’’*" y="*Y’’*"/>**

**<coordinate x="*X’’2*" y="*Y’’2*"/>**

**….**

**<coordinate x="*X’’p*" y="*Y’’p*"/>**

**</micrograph>**

**</particlepicking>**

Using dtd format (it can be found at <http://i2pc.cnb.csic.es/3dembenchmark/dtd/particlePicking.dtd>) it will be:

<?xml version="1.0" encoding="ISO-8859-1"?>

<!ELEMENT particlepicking (micrograph+)>

<!ELEMENT micrograph (coordinate\*)>

<!ATTLIST micrograph id NMTOKEN #REQUIRED>

<!ELEMENT coordinate EMPTY>

<!ATTLIST coordinate x NMTOKEN #REQUIRED>

<!ATTLIST coordinate y NMTOKEN #REQUIRED>

First and last line of the submitted document must be composed by the opening and closing clause of the element “particlepicking”. Next, write down the element “micrograph” with the attribute “id” showing the micrograph number. Inside the “micrograph” element insert the set of coordinates which has been picked on the micrograph. Each pair of coordinate will be specified by a “coordinate” element with two attributes “x” and “y”. The coordinate origin has been set in the upper left corner

“particle picking” clause at the beginning and end of the document and a “micrograph” clause for each micrograph are compulsory although no particle has been picked on the micrograph.

Find under an example of a document for a 3 micrographs dataset:

*<?xml version="1.0" encoding="ISO-8859-1"?>*

*<!DOCTYPE particlepicking PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN" "http://i2pc.cnb.csic.es/3dembenchmark/dtd/particlePicking.dtd">*

*<particlepicking>*

*<micrograph id="0001">*

*<coordinate x="1081" y="369"/>*

*<coordinate x="345" y="1693"/>*

*<coordinate x="972" y="122"/>*

*</micrograph>*

*<micrograph id="0002">*

*<coordinate x="1042" y="1098"/>*

*<coordinate x="222" y="716"/>*

*</micrograph>*

*<micrograph id="0003">*

*<coordinate x="976" y="345"/>*

*<coordinate x="1987" y="670”/>*

*<coordinate x="1076" y="108"/>*

*</micrograph>*

*</particlepicking>*