Data pre-processing program using machine learning

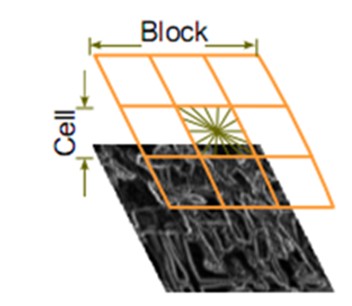
Xin He, Rice University

**[Abstract]** It is well known that not all data is available for cell tether force measurement experiments. The pick rate for a single molecule experiment which is similar to single cell is about 5%.Thus every time you analyze the data, you need to identify the valid data. So far, there is no quantitative selection criteria. The pre-processing of each experimental data relies on the researchers' artificial choices. Since everyone has their individual standard, it will eventually lead to different samples from the same experiment. To avoid the situation mentioned above, we developed a software that implements computer vision using toolkits from open source library. This software uses machine learning algorithms to let computers help researchers do data preprocessing. The software has now achieved an accuracy up to 85%.

**[Introduction]** The raw data of the experiment is too noisy and it is difficult to extract the traits from the values. Considering that people rely mainly on image features when discerning, we use the force curve image of raw data as the input sample. In order to extract image features, we use the HOG (Histogram of Oriented Gradients) algorithm to analyze the pattern of the image.

Using the HOG feature algorithm to assign a feature matrix to each raw data image, we can finally use a variety of machine learning classifiers to achieve the purpose of screening. Here we use the most commonly used SVM machine learning algorithm. In order to construct the classifier, we first provide two sets of artificially classified data as positive group and negative group to the program for training the classification model. Then we can use this classification model to effectively classify new data and filter out the useful data.

**[Histogram of Oriented Gradients]** The HOG algorithm is mainly used to find the local gradient direction and gradient intensity distribution of the image. Firstly, the input image needs to be normalized in the color space. Then we divide the image into several cells. The gradient direction histogram is calculated in each cell. All gradient directions are divided into 9 bins as the horizontal axis of the histogram while the gradient value corresponding to the angular range is added as the vertical axis. The features in each of these extracted cells are combined as a feature matrix of the entire image. This is so-called computer vision.



**(b)**

**(a)**

Figure 1. (a) The aim image is divided into several cells (b) The gradient direction of the cell is divided into nine direction blocks, and each pixel in the cell is weighted and projected in the histogram with a gradient direction (mapped to a fixed angular range).The gradient direction histogram is the 9-dimensional feature vector corresponding to the cell (because there are 9 bins).

**[**[**Support Vector Machines**](https://link.zhihu.com/?target=http%3A//en.wikipedia.org/wiki/Support_vector_machine)**]** The essence of SVM is actually a learning model to find the classification plane of the data points. The data points of two different groups are scattered in the data space and the algorithm tries to find a hyperplane to separate the two sets of data points and maximizes the distance between the hyperplane and the nearest data points (support vectors) on both sides. However, some data points are not well divided in the space of their own dimensions. Therefore, the kernel function theory can be used to map data points to higher dimensional space, which is more conducive to finding a suitable hyperplane for classification.

Figure 2. SVM try to find a hyperplane to separate the data points. The highlight points are so-called support vector which are used to determine the hyperplane.