R-CNN

Basic idea :

1) Takes an input image and generates category independent region proposals.They have used selective search method.
2) Each region proposal is warped into a fixed size as an input for the CNN. Regardless of the aspect ratio.
3) Each region proposal when passed through the CNN outputs a 4096 dimensional vector. This gives a typical feature matrix of dimensions 2000 x 4096.
4) The third module in this architecture is a class specific SVM , this SVM takes as an input each of the feature vectors of the 2000 regions and outputs classwise score of each of the regions. This makes the SVM matrix to be effectively 4096 X N dimensional where N is the number of classes. One advantage of this is that even if the number of classes is increased to a large number the matrix multiplication can still be carried out efficiently on a GPU..

More intricacies:

1) Region proposals with greater that 0.5 IoU overlap with the ground truth box are considered as positives for that particular class and negative for the rest. If the overlap is less than 0.5 then it is considered as negative / background. In each iteration 32 positive and 96 negative samples are made into a batch , the reason for this is because otherwise the number of negative samples will have an unfavourable bias.
2) It was observed that removing the fully connected features from the conv net produced quite good features , this showed that most of the representational ability of the network comes from the conv net and not the fully connected layers.
3) Infact a significant improvement is seen after fine tuning the fully connected layers . This shows that the conv part of the network learns general features while the fully connected part is responsible for task specific features.
4) Bounding box regression is also used for improving the predictions for localization , this improves the already proposed regions.
5) If we use CPMC for region proposals , warping the image to a fixed rectangular size cause the shape of the segment to be disregarded. A way for the network to learn even from the shape is to warp the image into a rectangular size but filling all the pixels apart from the region by the mean of the inputs so that after subtracting zero mean the shape of the region can also be considered by the network.