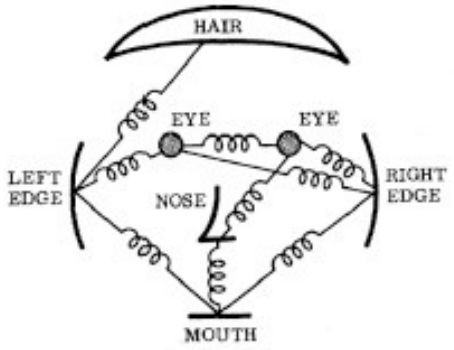
Object Detection

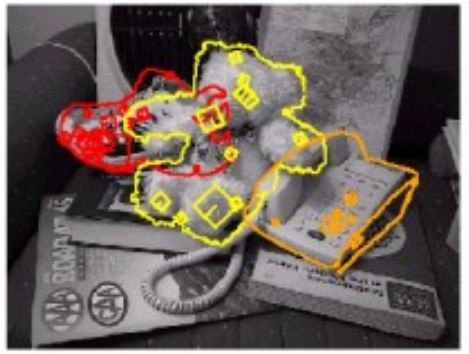
Image by kirkh.deviantart.com

Various kinds of recognition:

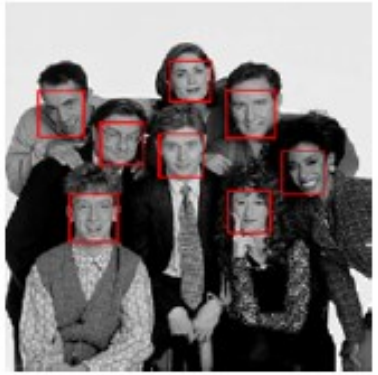
(a) Face recognition with pictorial structures
(b) Instance (known object) recognition
(c) Real-time face detection
(d) Feature-based recognition
(e) Instance segmentation using Mask R-CNN
(f) Pose estimation
(g) Panoptic segmentation
(h) Video action recognition
(i) Image captioning



(a)



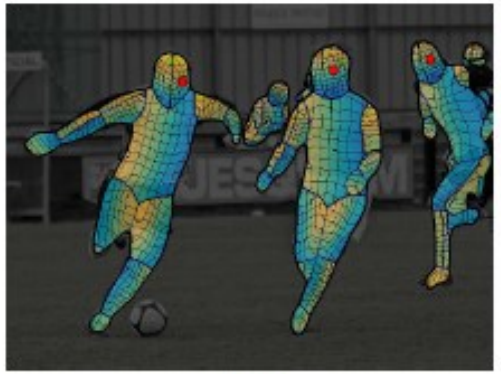
(b)



(c)

(d)

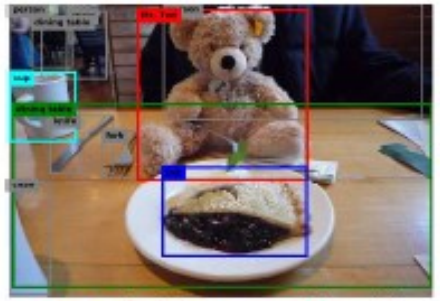
(e)



(f)

(g)

(h)



(i)

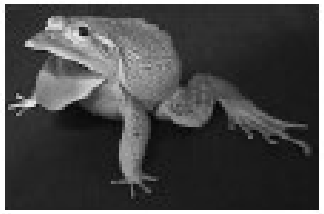
Object Recognition

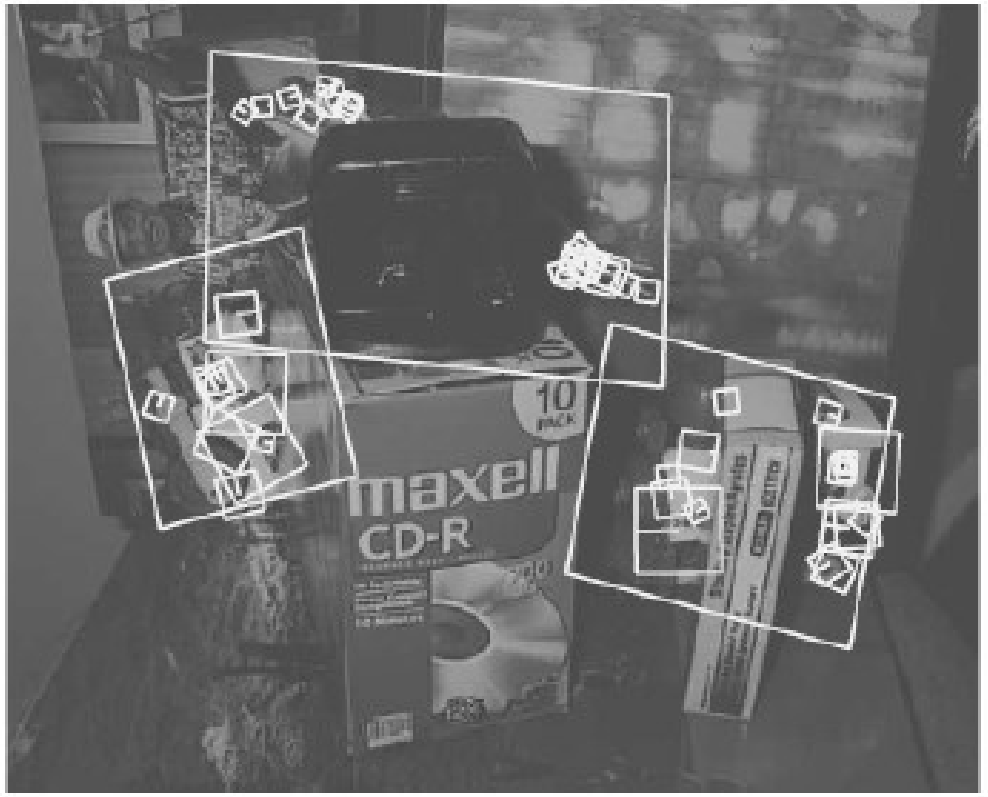
- General object recognition falls into two broad categories
- **Instance Recognition**: involves **re-recognizing a known 2D or 3D rigid object**, potentially being viewed from a novel viewpoint, against a cluttered background, and with partial occlusions
- **Class Recognition:** is also known as category-level or generic object recognition is the much more challenging problem of **recognizing any instance of a particular general class**, such as “cat”, “car”, or “bicycle”.

Instance Recognition: Geometric Alignment

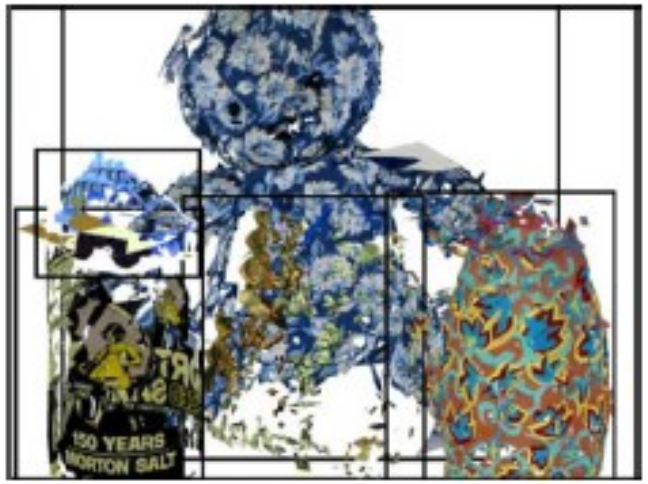
- To recognize one or more instances of some known objects, the recognition system
- Extracts a **set of interest points** in each database image and **stores** the associated descriptors (and original positions) in an indexing structure such as a search tree
- At recognition time, **features are extracted** from the new image and **compared** against the stored object features
- Whenever a **sufficient number of matching** features (say, three or more) are found for a given object, the system then invokes a **match verification stage**, to determine whether the **spatial arrangement of matching** features is consistent with those in the database image.
- Because images can be highly cluttered and similar features may belong to several objects, the original set of feature matches can have a large number of **outliers**.
  - **Hough transform** to accumulate votes for likely geometric transformations.

Instance Recognition: Geometric Alignment (cont.)





Instance Recognition: Geometric Alignment (cont.)



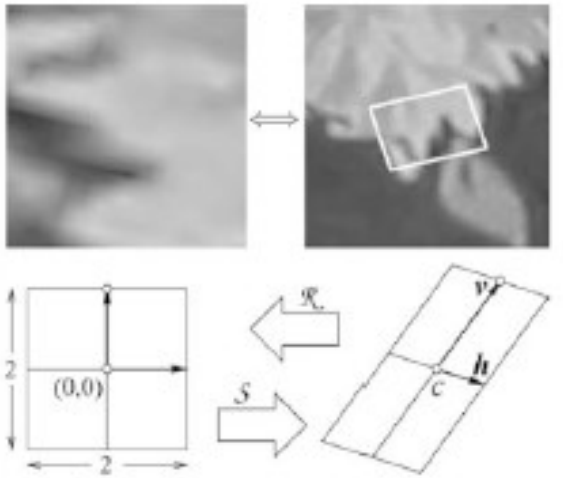


Image Classification: Feature-based methods

- **Bag of words** (also known as bag of features or bag of keypoints)
  - computes the distribution (histogram) of visual words found in the query image
  - compares this distribution to those found in the training images
- The biggest difference from instance recognition is the absence of a geometric verification stage since individual instances of generic visual categories, have relatively little spatial coherence to their features.

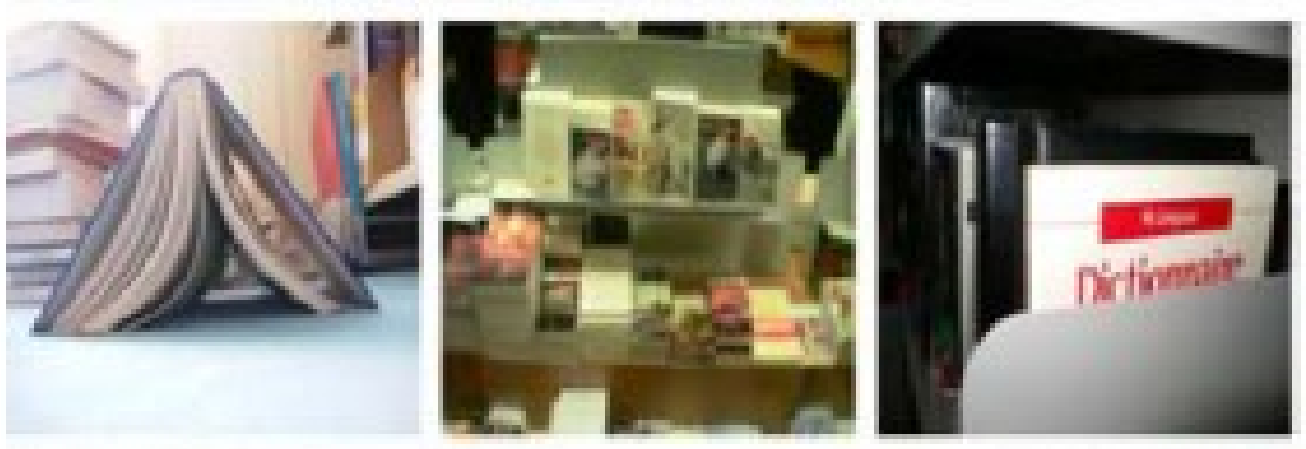


Image Classification: Feature-based methods (cont.)

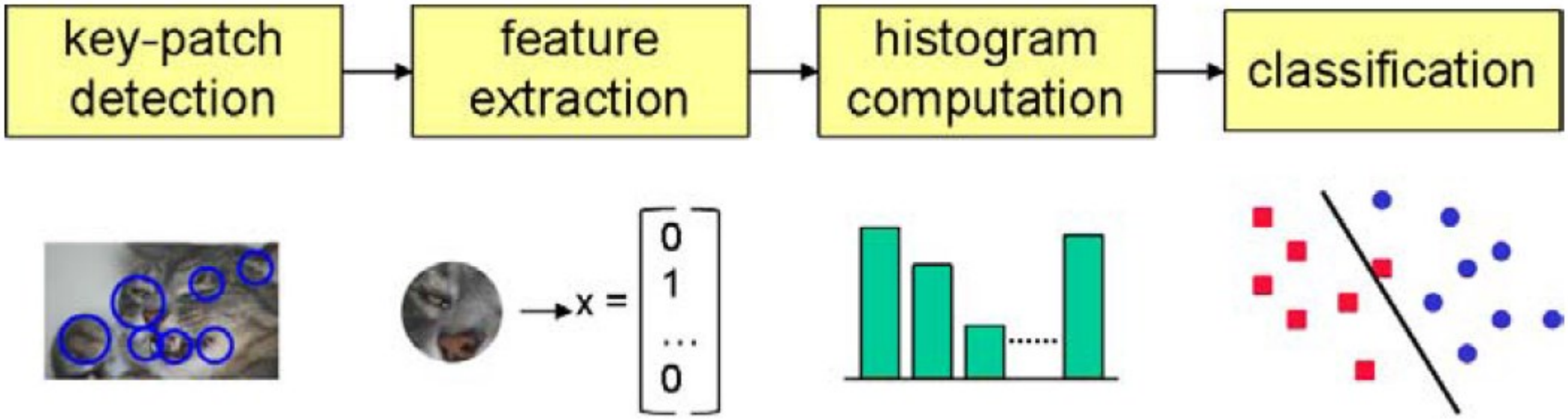


Image Classification: Feature-based methods (cont.)

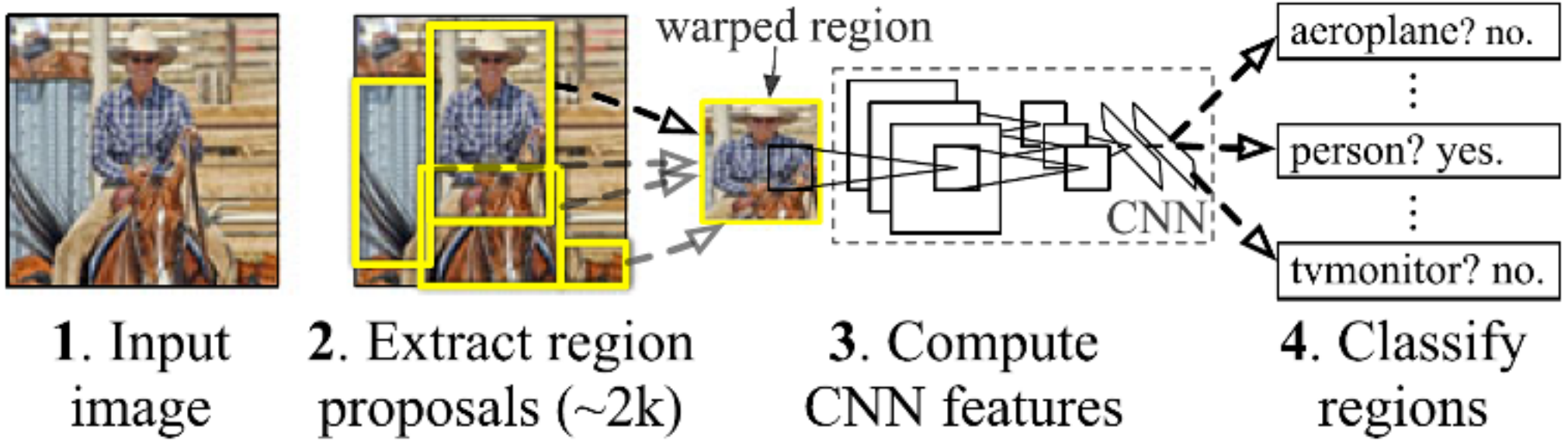
- Their original system used affine covariant regions and SIFT(Scale-invariant feature transform) descriptors, k-means visual vocabulary construction, and both a naive Bayesian classifier and support vector machines for classification.
- The debate about whether to use quantized feature descriptors or continuous descriptors and also whether to use sparse or dense features went on for many years.

Object Detection

- The main task in object detection is to put accurate bounding boxes around all the objects of interest and to correctly label such objects.

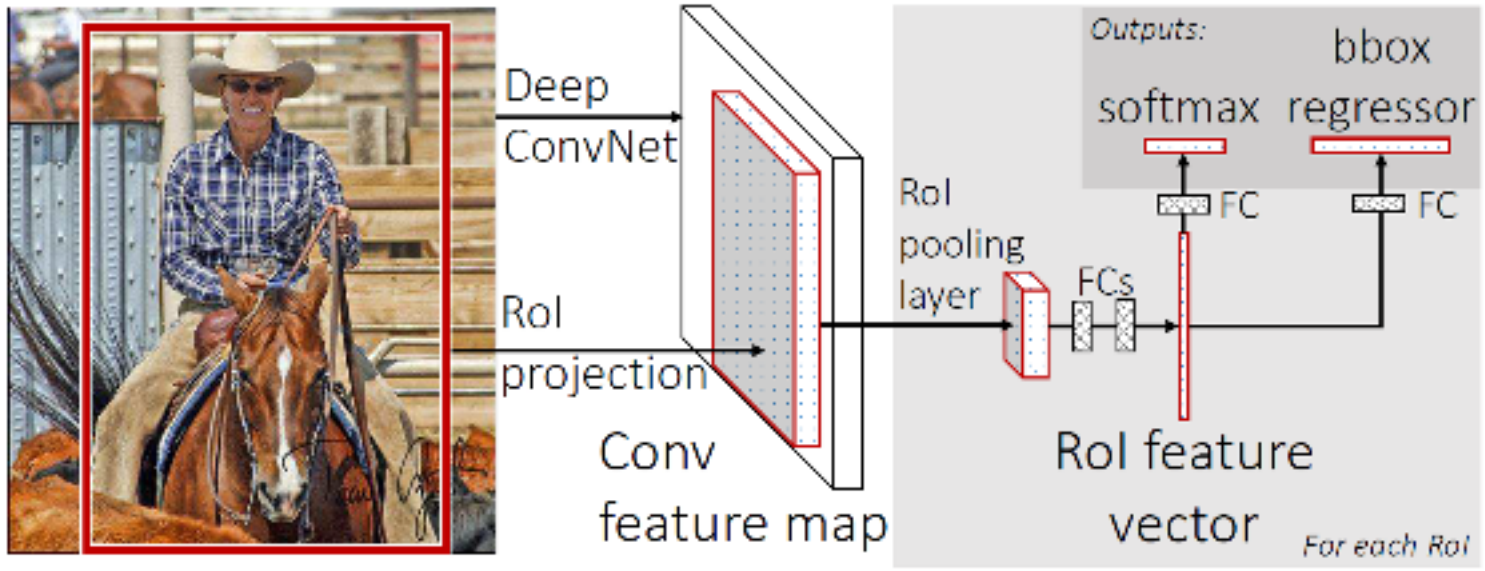
Object Detection: R-CNN

- One of the earliest object detectors based on neural networks is R-CNN, the Region-based Convolutional Network
  - This detector starts by extracting about 2,000 region proposals using the selective search algorithm. Each proposed regions is then rescaled (warped) to a 224 square image and passed through an AlexNet or VGG neural network with a support vector machine (SVM) final classifier.



Object Detection: Fast R-CNN

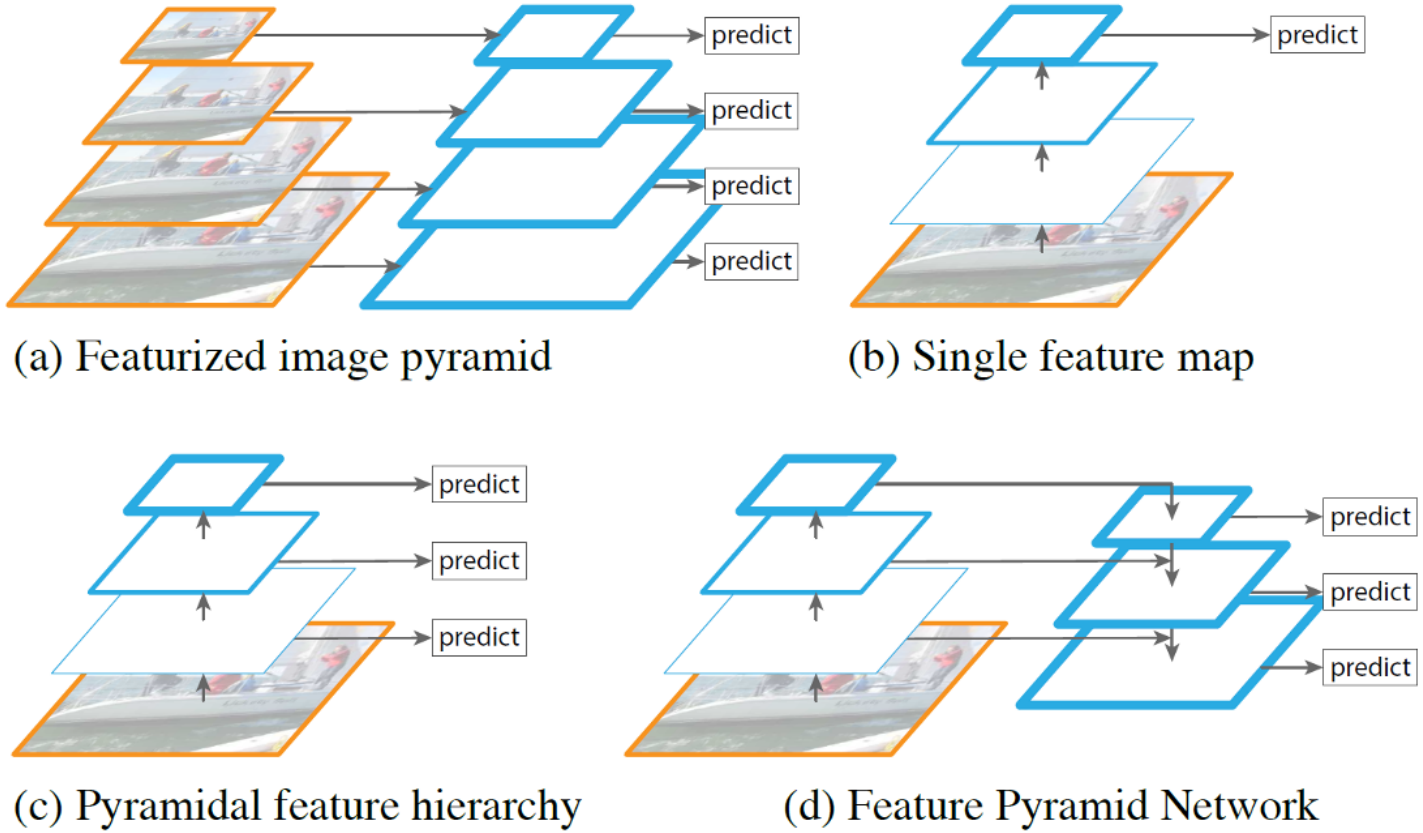
- Fast R-CNN interchanges the convolutional neural network and region extraction stages and replaces the SVM with some fully connected (FC) layers, which compute both an object class and a bounding box refinement



Object Detection: Faster R-CNN

- The Faster R-CNN system replaces the relatively slow selective search stage with a convolutional region proposal network (RPN), resulting in much faster inference.
- After computing convolutional features, the RPN suggests at each coarse location a number of potential anchor boxes, which vary in shape and size to accommodate different potential objects.
- R-CNN, Fast R-CNN, and Faster R-CNN all operate on a single resolution convolutional feature map . To obtain better scale invariance, it would be preferable to operate on a range of resolutions

Object Detection: Faster R-CNN (cont.)



(a) Featurized image pyramid

(b) Single feature map

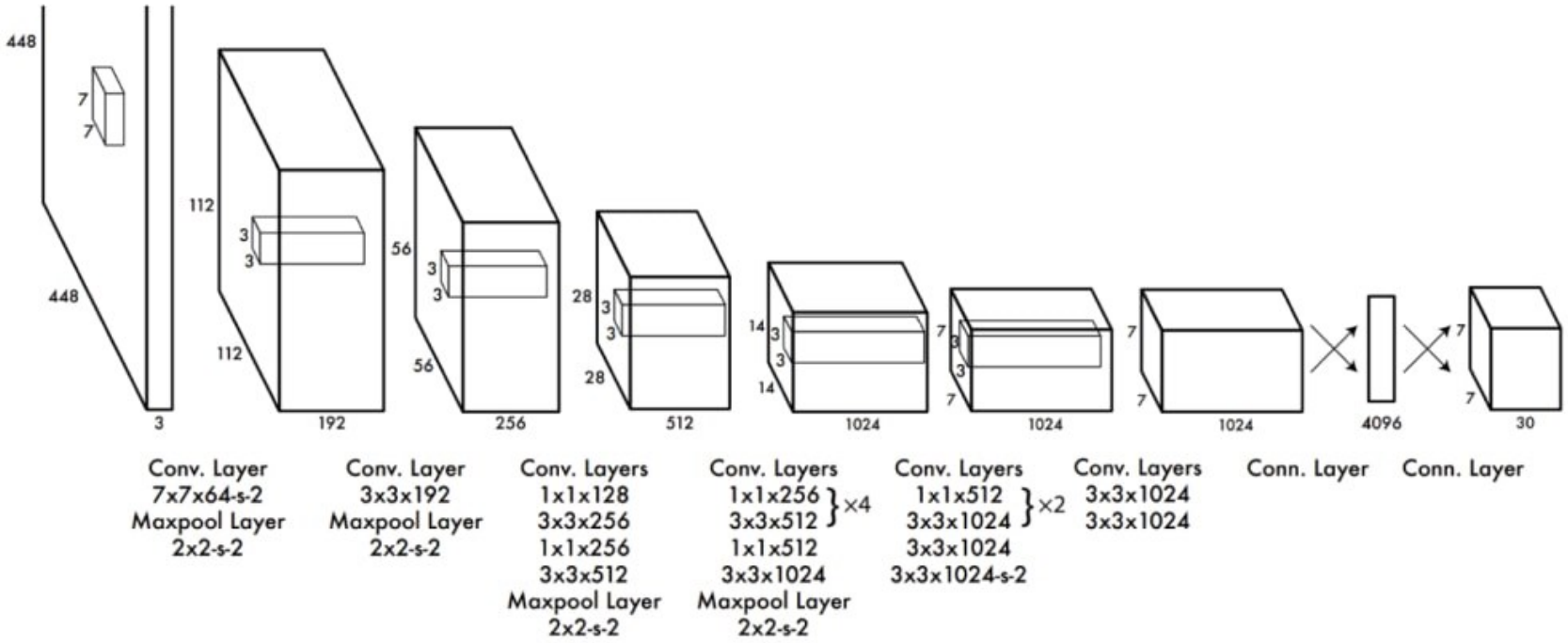
(c) Pyramidal feature hierarchy

(d) Feature Pyramid Network

Object Detection: Single-stage networks

- Single-stage network uses a single neural network to output detections at a variety of locations. Two examples of such detectors are SSD (Single Shot MultiBox Detector) and the family of YOLO (You Only Look Once) detectors

YOLO Neural Network: You Only Look Once



The Architecture. Our detection network has 24 convolutional layers followed by 2 fully connected layers. Alternating 1×1 convolutional layers reduce the features space from preceding layers. We pretrain the convolutional layers on the ImageNet classification task at half the resolution (224×224 input image) and then double the resolution for detection.

Thank You