International Journal of the Physical Sciences Vol. 6(29), pp. 6735-6740, 16 November, 2011
Available online at http://www.academicjournals.org/IJPS
DOI: 10.5897/IJPS11.1132
ISSN 1992 - 1950 © 2011 Academic Journals

Full Length Research Paper

A novel method for object detection based on graph theory

Shu Zhang* Mei Xie, Yuefei Zhang and Ting Wei

School of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu, China.

Accepted 10 August, 2011

Ideal object detection result in an image is an optimal free shape sub-window that tightly covers the object of interest. However, the sub-windows considered in widely-used sliding window method are limited to rectangles. This paper proposed a new graph-theoretic method which allowed the detection sub-window to be any shape for object detection. Firstly, local features responses were calculated by using locality-constrained linear coding (LLC) technique. Then the proposed method take advantage of local feature response and boundary information to construct an objective function for the whole image and global optimal solution is obtained by graph cut algorithm. We provided results on two challenging object detection datasets, and demonstrated that the proposed method can obtained better spatial support and higher detection precision than existing sliding window method.

Key words: Object detection, graph cut, sliding window, locality-constrained linear coding (LLC), support vector machine (SVM).

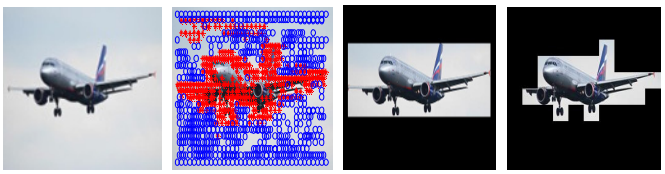
INTRODUCTION

People can easily know which and where object categories are present in an image, how about a computer? The development of the methods for automatic detection of the objects in images has been a challenge for computer vision and pattern recognition research. Object detectors determine whether a given object category is present in an image and estimate its spatial support. Many state-of-the-art approaches detect object using sliding window framework (Dalal and Triggs, 2005; Chum and Zisserman, 2007; Ferrari and Fevrier, 2008). Sliding window detection methods are well-suited for rigid objects with a fairly regular appearance pattern, and have been quite successful for detecting certain categories (Viola and Jones, 2001; Felzenszwalb et al., 2008). However, the high computational cost of evaluating the classifier (Addin et al., 2011) over a large set of windows is a severe limitation. The efficient sub-window search (ESS) algorithm of (Lampert et al., 2008) efficiently identifies the rectangle in the image that maximizes certain classifier functions. While these methods provide significant speed-ups over sliding window search, they are not suit for non-rectangular object.

In this paper, we propose a new graph-theoretic based object detection method which can obtain free-shape detection sub-window. Our method also extract local feature to represent object just like many sliding window method (Boureau et al., 2010; Ferrari et al., 2010). To address the problem of being sensitive to the feature noise, we additionally require the resulting sub-window to align well with boundary detected from the image. This way, the new localization objective is formulated as searching for a free-shape sub-window by striking a balance between two goals: (a) the optimal sub-window should cover as many positive-score features and as few negative-score features as possible, and (b) the sides of the optimal sub-window should have the maximum coincidence with the image boundary. In particular, we define a localization objective function which contains edge term and region term and show that the graph cut algorithm (Boykov and Jolly, 2001; Boykov and

*Corresponding author. E-mail: jlu_zhangshu@163.com.

Abbreviations: VOC, Visual object classes; **SVM,** support vector machine; **LLC,** locality-constrained linear coding; **BOF,** bag of feature; **SIFT,** scale-invariant feature transform; **SPM,** spatial pyramid matching; **HOG,** histograms of oriented gradients.



(a) (b) (c) (d)

Figure 1. Sliding window method versus proposed method; (a) original image, (b) image with local features mapped to a linear classifier's responses. Blue dots denote negative response of local feature; red dots denote positive response of local feature, (c) sliding window method gives imprecise detections (many background features) and (d) our method can more accurately detect object.

Table 1. A list of abbreviations used in this paper.

Abbreviations	Full meaning
SVM	Support vector machine
LLC	Locality-constrained linear coding
BOF	Bag of feature
SIFT	Scale-invariant feature transform
SPM	Spatial pyramid matching
HOG	Histograms of oriented gradients

Kolmogorov, 2004) can be used to find the optimal free-shape sub-window.

Figure 1 compares sliding window detection method with the proposed method; it is evident that our method can give a more precision location for object. One of the main contributions of this paper lies in constructing a new objective function for object detection and showing how to obtain this best-scoring region efficiently. Another contribution of this paper is that we give a new method (LLC-SVM) to calculate local feature response.

Approach

We first briefly overview the entire approach: (1) given training images in which the object of interest is marked by ground-truth rectangle, we train a linear support vector machine (SVM) classifier to distinguish that object category from any other. (2) Standard bag of feature (BOF)-SVM method has some disadvantages, in this paper we extend BOF to LLC feature and still use linear SVM to train object model. (3) Given a new image, we construct a uniform grid of patches over the image and extract local features associated with each patch. An objective function is constructed by the response of each patch and similarity between every adjacent patch. (4) We show that the problem of obtaining the global optimum solution of objective function is equivalent to the min-cut problem. In order to clearly describe the proposed method, a list of abbreviations used in this paper is shown in Table 1.

Train Classifiers

In this section, we use widely-used BOF-SVM method (Lampert et al., 2008) to train an object classifier which can map local feature to a response. In the BOF representation, a vocabulary of K visual words is obtained by clustering a sample of local features from the training images. Then BOF method represents an image as the frequency of visual words, in other words every image is transformed into K-dimensional histogram.
Using the histograms of the positive training examples (object) and the negative training examples (background), we learn a linear SVM model as shown in Equation 1

$$f(R) = \beta + \sum_i \alpha^i \left\langle h(R), h(R_i) \right\rangle \quad (1)$$

where i indexes the training examples, and α, β denote the learned weights and bias, and $h(R)$, $h(R_i)$ denote the histogram of image R and R_i respectively. Let $h^j(R)$ denote the count of the j-th visual word in the image R (the j-th bin of the histogram $h(R)$), then we rewrite the Equation 1 as below:

$$f(R) = \beta + \sum_{j=1}^{K} \omega^j h^j(R) = \beta + \sum_{i=1}^{N} \omega^{C_i} \quad (2)$$

Where $\omega^j = \sum_i \alpha^i h^j(R_i)$ is the j-th visual word's weight, and c_i is the index of the visual word that the i-th local feature maps to. Thus, the score of a region is the sum of its N features' word weights. The bias term β can be ignored for the purpose of maximizing $f(R)$. It is need to be emphasized that each local feature i is mapped to its closest visual word c_i, so its response is equal to c_i 's weight.

Although BOF-SVM is a wildly-used method for object detection, scene classification and object classification, it has two weak points if we want to map local feature to a response. First the standard BOF representation can achieve good classification performance only based on nonlinear classifier (Jianchao et al., 2009), but we need to obtain visual word's weight by using linear classifiers. This will increase the errors in local feature response. Second each local feature is mapped to its closest visual word, and this coarse quantification method will also produce many errors in local feature response. In this paper, we choose LLC algorithm which generalizes vector quantization to sparse coding followed by max pooling (Boykov and Kolmogorov, 2004), then a linear SVM can get good performance.

Locality-constrained linear coding (LLC)

LLC algorithm (Jinjun et al., 2010) includes two steps. Firstly, the input features are locally transformed into the representations that have some desirable properties such as compactness, sparseness. The code is obtained by decomposing the original feature on some codebook. Then the codes associated with local image features are pooled over the image.

Let X be a set of D-dimensional local descriptors extracted from an image, $X = [x_1, x_2, ..., x_N] \in R^{D \times N}$. Given a vocabulary of K visual words, $B = [b_1, b_2, ..., b_K] \in R^{D \times K}$, different coding schemes convert each descriptor into a K-dimensional code to generate the final image representation. This paper use LLC coding scheme.

Table 2. The whole training process for object model.

S/N	Extract training image's local features
1	Use k-mean clustering algorithm to obtain K visual words
2	Use LLC algorithm transform each local feature into K-dimensional sparse code
3	Every training image is represented by a K-dimensional feature using max pooling method, and then we obtain many positive and negative samples
4	Use linear SVM to obtain object model, namely every visual words' weight

A standard sparse coding problem is shown as eq.3, such a sparsity regularization term is selected to be the $l1$ norm of c_i ,

$$\arg\min_{C}\sum_{i=1}^{N}\| X_i - Bc_i \|^2 + \lambda \| c_i \|_{l1} \tag{3}$$

LLC incorporates locality constraint instead of the sparsity constraint in Equation 3. Specifically, the LLC code uses the following criteria:

$$\arg\min_{C}\sum_{i=1}^{N}\| X_i - Bc_i \|^2 + \lambda \| d_i \otimes c_i \|^2 \quad s.t.\ \ 1^T c_i = 1 \ \ \forall i \tag{4}$$

where $\otimes$ denotes the element-wise multiplication, and $d_i \in R^K$ is the locality adaptor that gives different freedom for each basis vector proportional to its similarity to the input descriptor. Specifically,

$$d_i = \exp(\frac{dist(x_i, B)}{\sigma}) \quad dist(x_i, B) = [dist(x_i, b_1), ..., dist(x_i, b_K)]^T \tag{5}$$

Note that the LLC code in Equation 4 is not sparse in the sense of $l0$ norm, but is sparse in the sense that the solution only has few significant values. We simply changes those small coefficients to be zero.

Clearly, we can obtain an approximated LLC fast encoding algorithm. The LLC solution only has a few significant values, or equivalently, solving eq.4 actually performs feature selection: it selects the local bases for each descriptor to form a local coordinate system. Instead of solving eq.4, we can simply use the M (M< D < K) nearest neighbours of x_i as the local bases B_i , and solve a much smaller linear system to get the codes:

$$\arg\min_{C}\sum_{i=1}^{N}\| X_i - c_i B_i \|^2 \quad s.t.\ \ 1^T c_i = 1, \ \ \forall i. \tag{6}$$

As K is usually very small, solving Equation 6 is very fast. When every local feature transforms into K-dimensional sparse code, max pooling method has been used in whole image. Finally, image is represented by a K-dimensional feature. The training process for object model is shown in Table 2.

Objective function

Given a new image, we compute local feature response in two steps: first every local feature is transformed into K-dimensional sparse code using LLC algorithm, then the local feature response is equal to the inner product between K-dimensional sparse code and K-dimensional visual word's weight. Figure 1 (b) shows the result of local feature response, blue dots denote negative response of local feature; red dots denote positive response of local feature.

We construct a uniform grid of patches over image. Let $A = (A_1, ..., A_k, ..., A_N)$ be a binary vector whose components A_k specify assignments to patch k, N is the number of patches over image. Each A_k can be object or background, so the indication vector A give results of detection and coarse segmentation. Then, we obtain objective function by imposing local feature response and boundary information:

$$E(A) = \mu \times R(A) + B(A) \tag{7}$$

where μ is relative importance coefficient, $R(A)$ is regional term, $B(A)$ is edge term

$$R(A) = \sum_{k=1}^{N} R_k(A_k)$$

$$B(A) = \sum_{\{p,q\} \in C} B_{\{p,q\}} \times \delta(A_p, A_q)$$

and $\delta(A_p, A_q) = 1$ if $A_p \neq A_q$, $\delta(A_p, A_q) = 0$ if $A_p = A_q$. C is the set of all adjacent patches over image.

$R_k(A_k)$ can be calculated by local feature response. First, the response $res(k)$ assigned to a patch is the sum of the response for all local features located within that patch. Then we use logistic sigmoid function to maps the response into [0-1] interval, so we obtain the probability of a patch belongs to object. $R_k(A_k)$ is calculated as below:

$$R_k(A_k = 1) = 1 - 1/(1 + \exp(-res(k) \times \rho)) \quad R_k(A_k = 0) = 1/(1 + \exp(-res(k) \times \rho)) \tag{8}$$

This paper defines $A_k = 1$, when the k-th patch belongs to object, and $A_k = 0$ when the k-th patch belongs to background. ρ is a scale factor, in this paper we set 2. The regional term

Table 3. Each edge is assigned a nonnegative weight.

Edge	Weight
n-links {p,q}	$B_{\{p,q\}}$
t-links {k,S}	$\mu \times R_k(A_k = 0)$
t-links {k,T}	$\mu \times R_k(A_k = 1)$

Table 4. Value of parameters in object detection system.

Parameters	Value
Num SIFT center	500
KNN in LLC	5
Importance coefficient	$\mu = 2$
Scale factor	$\rho = 2$

$R(A)$ represents the sum of penalty for assigning every patch to object and background.

$B_{\{p,q\}}$ is calculated by boundary property. For color image, first each color channel is quantified to have 8 different values, and each patch can be represented by 24-dimentional histogram. Then, the similarity of two histograms is calculated by histogram intersection function (Swain and Ballard, 1991), see Equation 9.

$$B_{\{p,q\}} = \sum_{i=1}^{D} \min(hist_p(i), hist_q(i)) \quad (9)$$

where $hist_p(i)$ represents the i-th component of the p-th patch's histogram. For gray image, the only difference is that we quantize gray scale to have 8 different values, and each patch can be represented by 8-dimentional histogram. It is deserved to be mentioned that boundary information is introduced into objective function by edge term.

It is clear that the objective function will be small value when we obtain a good spatial support for object, so the final goal is to obtain a global optimal solution of objective function (Equation 10).

$$A_{opt} = \arg\min_A E(A) = \arg\min_A (\mu \times R(A) + B(A)) \quad (10)$$

Graph cut

Now we transform our objective function into graph-theoretic structure. An undirected graph $G = \langle V, \varepsilon \rangle$, is defined as a set of nodes (vertices V) and a set of undirected edges (ε) that connect these nodes. In this paper, each patch is considered as a node and every adjacent patch have an undirected edge. There are two additional nodes called terminals: source terminal S represent object and sink terminal T represent background. Furthermore, each patch will connect these two terminals. We define that the set of edges ε consists of two types of undirected edges: n-links (neighbourhood links) and t-links (terminal links). Each edge $e \in \varepsilon$ in the graph is assigned a nonnegative weight W_e. In this paper, t-links is assigned corresponding regional term and n-links is assigned corresponding edge term, which is shown in Table 3. In graph theory, a cut is a subset of edges $C \subset \varepsilon$ such that the terminals become separated on the induced graph $G(C) = \langle V, \varepsilon \setminus C \rangle$. It is normal to define the cost of a graph cut as the sum of the weight of the edges that it severs. It is clear that a minimum cost cut in graph gives optimal solution for Equation 10. A fast implementation of min-cut algorithms can be an issue. The most straight-forward implementations of the standard graph cut algorithms, for example, max-flow (Ford and Fulkerson, 1962), can be slow. Boykov and Kolmogorov (2004) describes a new version of the max-flow algorithm that (on typical in vision examples) significantly outperformed the standard techniques. This paper also uses this new graph cut algorithm to obtain optimal solution.

Experiments

Object detection database

The experiments used UIUC car database and PASCAL visual Object classes (VOC) 2010 database. UIUC car database contains images of side views of cars for use in evaluating object detection algorithm. It has 1050 training images (550 car and 500 non-car images), the size of each training image is 100×40. In addition, this database has 170 test images, containing 200 cars at roughly the same scale as in the training images. They are of different resolutions and include instances of partially occluded cars, cars that have low contrast with the background, and images with highly textured backgrounds.

PASCAL VOC 2007 database contains a total of 9963 annotated images from 20 object classes. The images span the full range of consumer photographs, including indoor and outdoor scenes, close-ups and landscapes, and strange viewpoints. The dataset is extremely challenging due to the wide variety of object appearances and poses and the high frequency of major occlusions.

UIUC car database results

The whole object detection system contains many important details and parameters which are determined through experiments. We extract dense scale-invariant feature transform (SIFT) at three scales (16, 24, 32 pixels) with grid spacing of 8 pixels, and uniformly divide every image into 10×10 patches, the other parameters see Table 4. The visual detection results are shown in Figure 2. The detection sub-windows based on our method tightly cover the object of interest, even though the scale of object is unknown (Figure 2).

To evaluate the detection results of the proposed method, this paper proposed a simple fusion method. First we retained all 100×40 windows which cover at least 80% object spatial support, and then the response of each window is calculated by summing the response of all local features located within that window. The final detection window can be obtained using non-Maxima suppression. We adopt an appropriate evaluation criterion that a location output by the detector to be evaluated as a correct detection, if its center lie within a

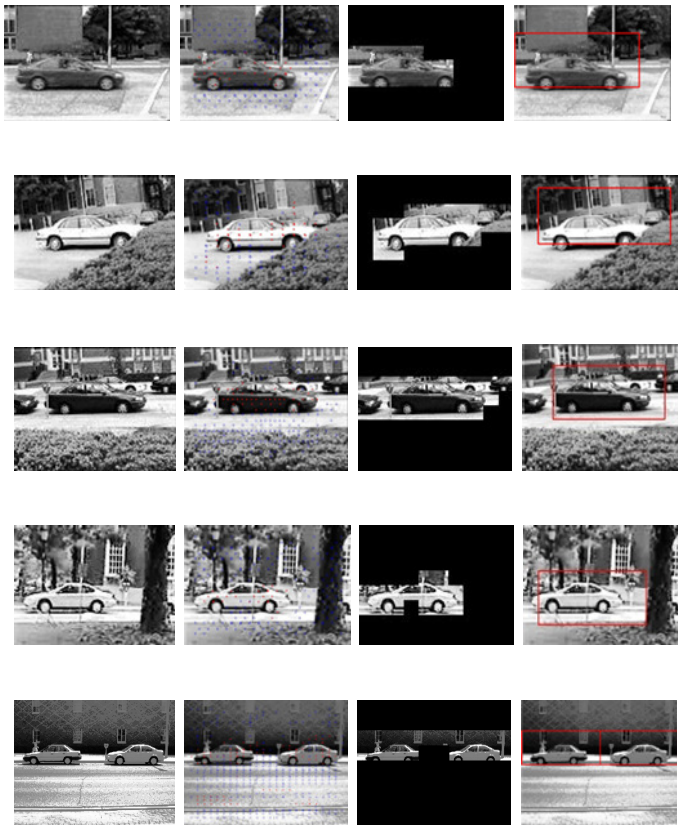


Figure 2. Detection results in UIUC database. The first column is original images, the second column is results of local feature response, the third column is detection results using our proposed method, and the last column is final detection results based on simple fusion method.

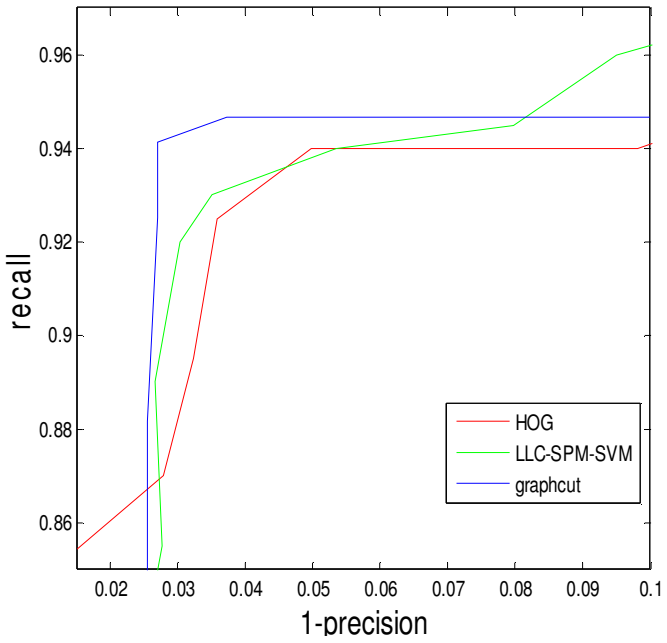


Figure 3. Car detection recall-precision curves for three different object detection method.

rectangle of a 50×20 size centered at the ground-truth location.

As baselines for comparison, this experiment implemented two other sliding window methods. The first one is histograms of oriented gradients (HOG) (Dalal and Triggs, 2005) which is a state of the art method for object detection. It achieves very nice detection results and is widely used. The second one is LLC-spatial pyramid matching (SPM)-SVM method (Jinjun et al., 2010). The comparison result is shown in Figure 3. It is evident that the proposed detection method has higher precision and recall rate than the other two methods. Especially, when precision is 96%, the recall rate of the proposed method is close to 95%, but LLC-SPM-SVM method and HOG method is about 93%.

PASCAL visual object classes (VOC) 2007 results

In order to prove that the proposed method can obtain good detection results for non-rectangle object, we choose 5 animal classes including horse, cat, cow, dog, and sheep for testing. In this experiment, we adopted the same parameters previously mentioned. In addition, the training and validation images were used for constructing the visual words and deriving the feature scores, and the test images were used for testing the performance of object detection.

As in many previous works (Chum and Zisserman, 2007; Harzallah et al., 2009; Senjian et al., 2009), the relative overlap between the detection sub-window C and the manually labelled ground-truth sub-window C_{gt} on the test images is usually used to measure the localization accuracy:

$$S(C,C_{gt})=\frac{C\cap C_{gt}}{C\cup C_{gt}} \qquad (11)$$

A detection result C is regarded to be correct if $S(C,C_{gt})\geq 0.5$. In VOC 2007 database, the ground truth C_{gt} in an image is a rectangle (or multiple rectangles when multiple objects are present) around the object of interest. Table 5 gives the correct localization rate of the proposed method and the ESS algorithm (Lampert et al., 2008). These results show that the proposed method has a higher correct localization rate than the ESS algorithm. Visual detection results show that, the proposed method can obtain a good special support, even for non-rectangle object (Figure 4).

Conclusions

We introduced an efficient graph cut based object detection. Different from previous sliding window search based object detection methods, we considered both

Table 5. The correct localization rate of the proposed algorithm and the ESS algorithm on VOC 2007 database.

Dataset	Proposed	ESS
Horse	0.441	0.388
Cat	0.489	0.422
Cow	0.251	0.176
Dog	0.407	0.389
Sheep	0.208	0.095

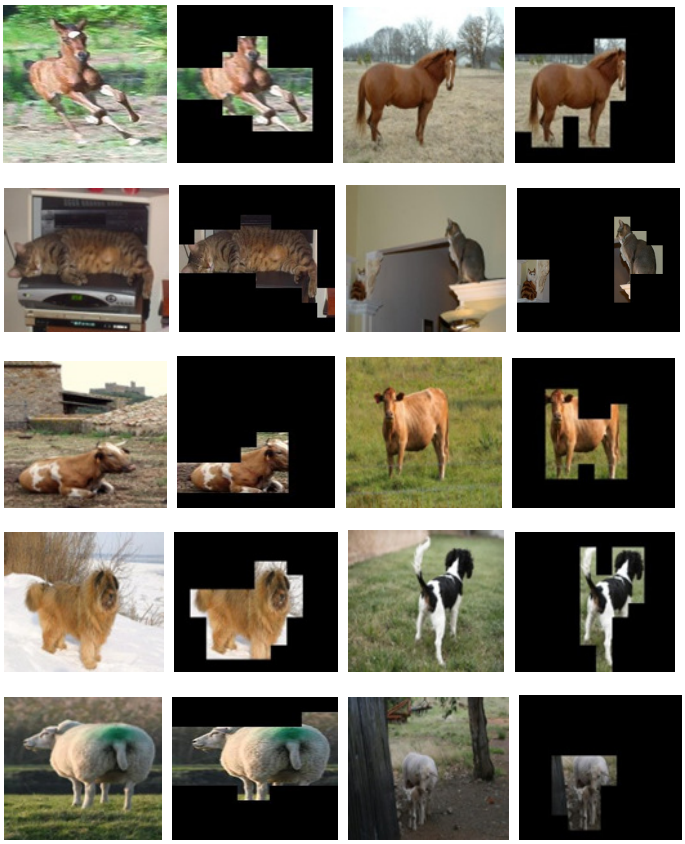


Figure 4. Object detection results using the proposed algorithm on VOC 2007 dataset.

object features and boundary information for object localization. Then we demonstrated that the proposed method has three advantages over some state of the art methods in two challenging datasets. First our method can obtain better spatial support, especially for non-rectangle object. Second our method can obtain higher detection precision. Third our method does not take into account object scale which is a big obstacle for sliding window Model.

Future work

There are some issues remain to be researched, and future work could include the following improvements: 1) use other local feature and adopt different method to calculate local feature response. 2) change smallest spatial tokens (for example, pixel or superpixel (Hoiem et al., 2005) replace patch) and improve objective function; the final goal is to achieve object detection and object segmentation simultaneously.

REFERENCES

Addin O, Sapuan SM, Othman M, Ahmed Ali BA (2011). Comparison of Naive bayes classifier with back propagation neural network classifier based on f - folds feature extraction algorithm for ball bearing fault diagnostic system. Int. J. Phy. Sci., 6(13): 3181-3188.

Boureau YL, Bach F, LeCun Y, Ponce J (2010). Learning mid-level features for recognition. Computer Vision and Pattern Recognition (CVPR).

Boykov Y, Kolmogorov V (2004). An experimental comparison of min-cut/max- flow algorithms for energy minimization in vision. IEEE. Transactions on Pattern Analysis and Machine Intelligence., 26(9): 1124-1137.

Boykov Y, Jolly MP (2001). Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images. IEEE International Conference on Computer Vision (ICCV).

Chum O, Zisserman A (2007). An Exemplar Model for Learning Object Classes. Computer Vision and Pattern Recognition(CVPR).

Dalal N, Triggs B (2005). Histograms of oriented gradients for human detection. Computer Vision and Pattern Recognition (CVPR).

Felzenszwalb P, McAllester D, Ramanan D (2008). A discriminatively trained, multiscale, deformable part model. Computer Vision and Pattern Recognition (CVPR)

Ferrari V, Fevrier L (2008). Groups of Adjacent Contour Segments for Object Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence., 30(1): 36-51.

Ferrari V, Jurie F, Schmid C (2010). From Images to Shape Models for Object Detection. IJCV. 87(3): 284-303.

Ford L, Fulkerson D (1962). Flows in Networks. Princeton University Press.

Harzallah H, Jurie F, Harzallah C (2009). Combining efficient object localization and image classification. IEEE International Conference on Computer Vision (ICCV).

Hoiem D, Efros AA, Hebert M (2005). Geometric context from a single image. IEEE International Conference on Computer Vision (ICCV).

Jianchao Y, Kai Y, Yihong G, Huang T (2009). Linear spatial pyramid matching using sparse coding for image classification. Computer Vision and Pattern Recognition (CVPR)

Jinjun W, Jianchao Y, Kai Y, Fengjun L (2010). Locality-constrained Linear Coding for image classification. Computer Vision and Pattern Recognition (CVPR).

Lampert CH, Blaschko MB, Hofmann T (2008). Beyond sliding windows: Object localization by efficient subwindow search. Computer Vision and Pattern Recognition (CVPR)

Senjian A, Peursum P, Senjian L, Venkatesh S (2009). Efficient algorithms for subwindow search in object detection and localization. Computer Vision and Pattern Recognition (CVPR)

Swain M, Ballard D (1991). Color indexing. IJCV., 7(1): 11–32.

Viola P, Jones M (2001). Rapid object detection using a boosted cascade of simple features. Computer Vision and Pattern Recognition (CVPR).