Pattern Recognition 48 (2015) 3542–3559

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Feature representation for statistical-learning-based object detection: A review ☆

Yali Li [a,b,c], Shengjin Wang [a,b,c,*], Qi Tian [d], Xiaoqing Ding [a,b,c]

[a] State Key Laboratory of Intelligent Technology and Systems, PR China
[b] Tsinghua National Laboratory for Information Science and Technology, PR China
[c] Department of Electronic Engineering,Tsinghua University,Beijing 100084, PR China
[d] Department of Computer Science at the University of Texas at San Antonio (UTSA), United States

ARTICLE INFO

Article history:
Received 11 June 2014
Received in revised form
12 February 2015
Accepted 15 April 2015
Available online 24 April 2015

Keywords:
Object detection
Feature representation
Feature learning
Dimension reduction
Feature combination

ABSTRACT

Statistical-learning-based object detection is an important topic in computer vision. It learns visual representation from annotated exemplars to identify semantic defined objects in images. High-performance object detection is usually carried out in feature space and effective feature representation can improve the performance significantly. Feature representation is the encoding process which maps raw image pixels inside local regions into discriminant feature space. The motivation of this paper is to present a review on feature representation in recent object detection methods. Visual features applied in object detection are categorized according to the differences in computation and visual properties. The most valued features are introduced and discussed in detail. Representative extensions are introduced briefly for comparison. Descriptive power, robustness, compactness as well as computational efficiency are viewed as important properties. According to these properties, discussions are presented on the advantages and drawbacks of features. Besides, generic techniques such as dimension reduction and combination are introduced. Through this review, we would like to draw the feature sketch and provide new insights for feature utilization, in order to tackle future challenges of object detection.

1. Introduction

Object detection is a fundamental and important problem in computer vision. The goal of object detection is to find and identify the objects with semantic definitions in digital images. Object detection is the cornerstone of computer vision based applications, such as robotics and video analysis. Nowadays, statistical-learning-based object detection which learns representation of objects from annotated exemplars has been widely researched. It has displayed effectiveness and efficiency especially in cluttered environments. The general framework of statistical-learning-based object detection is illustrated in Fig. 1. In the training stage, objects and background sample images are collected. Through a feature representation, images are mapped into discriminant feature space. Statistical learning is then applied to learning the object representation models. In the detection stage, test images are firstly mapped into the same feature space. The learnt object representation models are used to infer whether objects exist and to obtain the locations of the objects. In general, the feature representation, the learning/inference algorithms and the training samples are three factors that affect object detection performance.

This review is mainly for feature representation methods in statistical-learning-based object detection. Feature representation is the mapping from raw image pixels to a discriminant high-dimensional data space. It bridges low-level pixels with the input of learning/inference algorithms. Since object detection is challenged by the sematic gap between low-level pixels and high-level semantic definitions, the feature representation is quite important to construct high-performance object detection systems. Although there is a survey on object recognition [1], we aim to present detailed introduction and comprehensive discussions on visual features. We mainly concentrate on local feature representation methods in object detection. The most valuable features are chosen to be introduced in detail. Typical generalizations are briefly introduced. Besides, we view descriptive power, robustness, compactness, computational efficiency as valued properties for the feature representation. The discussions and comparisons are presented on these aspects. The motivation of this review includes the following: (1) *to provide a reference for researchers by categorizing feature representation methods according to computing procedure*

☆This document is a collaborative effort. The work is supported by the National High Technology Research and Development Program of China (863 program) under Grant no. 2012AA011004 and the National Science and Technology Support Program under Grant no. 2013BAK02B04.

* Corresponding author.
E-mail addresses: liyali@ocrserv.ee.tsinghua.edu.cn (Y. Li), wsj@ocrserv.ee.tsinghua.edu.cn (S. Wang), qitian@cs.utsa.edu (Q. Tian), dxq@ocrserv.ee.tsinghua.edu.cn (X. Ding).

http://dx.doi.org/10.1016/j.patcog.2015.04.018
0031-3203/

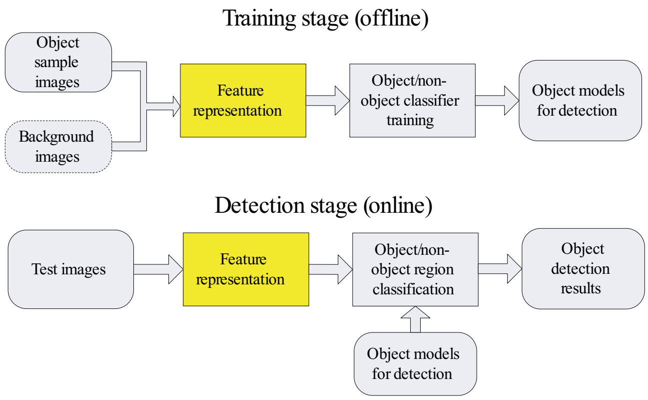

Fig. 1. The framework of statistical learning based object detection. This review mainly focuses on the feature representation components of this framework.

and visual properties; (2) *to draw the development procedure by introducing features from each category in chronological order;* (3) *to shed new insights into feature design and application for object detection by presenting the comparison of different visual features.*

The remainder part of the paper is organized as follows. In Section 2, we present the classification of features in statistical-learning-based object detection. Section 3 introduces and compares existing feature engineering methods. Section 4 introduces the feature learning methods. Generic techniques in feature representation are presented in Section 5. Evaluation is discussed in Section 6, and a summary is presented in Section 7.

2. Classification of features in object detection

A large amount of feature representation methods are integrated into statistical-learning-based object detection. In this review, we mainly focus on local features rather than global features which rely on subspace learning for whole image patches. We briefly divide the features into two classes as *human-engineering-based features* and *learning-based features*. The engineering based features are further divided into four classes as *gradient features*, *pattern features*, *shape features* and *gray-tone/color features*. The typical features belonging to different classes are listed in Table 1. We categorize the features based on the differences in feature computation and visual properties. *Gradient features* are based on constructing the histograms from gradient filtering outputs, including the popular SIFT [2] and HOG [11]. *Pattern features* refer to those that analyze relations of neighboring pixels or subregions for local image representation, including Gabor [30], LBP [31] and Haar-like features [38]. *Shape features* mainly focus on shape description and are based on pre-detected contour fragments, including shape context [47]. *Gray-tone/color features* mainly focus on the probabilistic representation constructed in intensity/color spaces, including Entropy saliency [61], CSS [66], color SIFT [73] and color names [74]. In contrast to the above features involving much human engineering, *learning-based features* rely on learning algorithms to automatically obtain feature representation from local images, including DeCAF [77] and the feature representations learned by deep neural networks. We present the computation details of the most valuable features from each category in Section 3. Based on the introduction and comparison of different features, we also present discussions on the descriptive power, invariance properties, computational efficiency and compactness on existing feature representation methods.

3. Human-engineering-based feature representations

In this section, we mainly introduce the hand-crafted feature representations with human engineering and discuss the detailed feature computation in object detection. The features are classified by the differences in computation and visual properties. Besides of the introduction of typical features and the generalizations, we also discuss the advantages and drawbacks.

3.1. Gradient features

Gradient information is quite useful in image interpretation. Gradient features represent objects by the distributions of gradient intensities and orientations over spatial regions. Typical gradient features including the widely applied SIFT and HOG have been playing an important part in object detection.

3.1.1. Scale-invariant feature transform (SIFT)

Scale-invariant feature transform (SIFT) [2][1,2] describes objects by gradient information around identified keypoints. Sparse defined SIFT is the combination of DoG keypoint detection and histogram-based gradient representation. It has four steps, scale-space extrema searching, sub-pixel keypoint refining, dominant orientation assigning and description (Fig. 2). Firstly, DoG (difference of Gaussian) scale spaces are constructed by obtaining the difference of two nearby Gaussian smoothed layers, as

$$D(x,y,\sigma) = (G(x,y,k\sigma) - G(x,y,\sigma)) * I(x,y). \tag{1}$$

where k is a multiplicative factor. $G(x,y,\sigma)$ is the Gaussian scale-space kernel, which is given as

$$G(x,y,\sigma) = \frac{1}{2\pi\sigma^2} \exp\left(-(x^2+y^2)/2\sigma^2\right). \tag{2}$$

Potential keypoints are identified by exhaustively searching for the extremas in scale spaces (Fig. 3). Sub-pixel accuracy of keypoint localization is further achieved by interpolation with Taylor expansion of $D(x,y,\sigma)$. The DoG function values at the extremas are used to reject low contrast potentials and Hessian matrix $\mathbf{H}$ is used to eliminate the potentials with strong edge responses at only one orientation. The located keypoints are denoted by location (x,y) and scale σ.

Local gradient information around identified keypoints is used for SIFT feature description. Suppose the Gaussian smoothed image with fixed scale σ is $L(x,y) = G(x,y,\sigma) * I(x,y)$. The gradient magnitude and orientation are computed as

$$m(x,y) = \sqrt{L^2(\Delta x,y) + L^2(x,\Delta y)},\ \theta(x,y) = \tan^{-1}(L(x,\Delta y)/L(\Delta x,y)). \tag{3}$$

Dominant orientations are assigned for keypoints by finding peaks of weighted gradient orientation histogram. Each point inside the local window is casted into the histogram by discretized orientations with the weight determined by gradient magnitude and Gaussian window function. Finally, the feature descriptor is the concatenation of orientation histogram in divided subregions. The local region determined by (x,y,σ) is divided into 4×4 subregions and weighted orientation histograms with 8 bins are computed from the divided subregions. The final 128 feature descriptor is normalized to unit length to reduce the influences of illumination changes. Rotation invariance is achieved by normalization with dominant orientation. The whole SIFT feature is formed from the location x,y, the scale σ, the orientation θ and associated gradient histogram descriptors.

In addition to the standard SIFT description, there are dense SIFT descriptors computed in uniformly and densely spaced local regions. SIFT features are highly distinctive with scale and rotation invariance. They are several extensions of SIFT features as follows:

[1] http://www.cs.ubc.ca/lowe/keypoints/
[2] http://www.robots.ox.ac.uk/vedaldi/code/sift.html/