J. Vis. Commun. Image R. 25 (2014) 1251–1261

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Hierarchical Implicit Shape Modeling

Parvin Razzaghi [a,*], Shadrokh Samavi [a,b]

[a] Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan, Iran
[b] Department of Electrical and Computer Engineering, McMaster University, Hamilton, Canada

ARTICLE INFO

Article history:
Received 20 March 2013
Accepted 29 December 2013
Available online 8 January 2014

Keywords:
Object recognition
Statistical part-based object recognition
Implicit shape model
Hierarchical Implicit Shape Model
Hierarchical star graph
Discriminative parts
Parts filter
Histogram of gradients

ABSTRACT

In this paper, a new hierarchical approach for object detection is proposed. Object detection methods based on Implicit Shape Model (ISM) efficiently handle deformable objects, occlusions and clutters. The structure of each object in ISM is defined by a spring like graph. We introduce hierarchical ISM in which structure of each object is defined by a hierarchical star graph. Hierarchical ISM has two layers. In the first layer, a set of local ISMs are used to model object parts. In the second layer, structure of parts with respect to the object center is modeled by global ISM. In the proposed approach, the obtained parts for each object category have high discriminative ability. Therefore, our approach does not require a verification stage. We applied the proposed approach to some datasets and compared the performance of our algorithm to comparable methods. The results show that our method has a superior performance.

1. Introduction

Object recognition is one of the most important research areas in computer vision. Object recognition has received much interest in recent decades. Many computer vision systems can benefit from an accurate object recognition stage. Some computer vision systems include scene understanding, video surveillance, and human robot interactions, etc. Change within class object appearance, viewpoint, scale and illumination pose major challenges to the object recognition task.

Up to now, many approaches are introduced in the field of object recognition. Some of these approaches are part-based modelings in which it is assumed that each object consists of some parts which are placed in a special structure. In statistical part-based modeling, structure of parts is represented by a graph. Statistical part-based modeling is divided into three categories based on type of the structure [1].

The first statistical part-based modeling is Bag of visual Words (BoW) in which it is assumed that all parts are independent. Hence, the learning and inference procedures based on BoW are easy to implement. In these approaches, at first, interest points are extracted and described. These features are then quantized into visual words using a clustering algorithm. Eventually, each object is represented by a histogram of visual words [2]. The BoW models are extended to topic modeling [3]. In topic modeling, each object is modeled as a mixture of topics. Each topic is a probability distribution over visual words which frequently occur together. Hence, the final representation for a particular object is being composed of the mixture of the histograms corresponding to each topic. Two important algorithms based on the topic modeling are *probabilistic Latent Semantic Analysis* (pLSA) [4] and *Latent Dirichlet Allocation* (LDA) [5]. In LDA, the topic distribution has a Dirichlet prior, whereas in pLSA the topic distribution is uniform. It should be noted that, topic modeling is based on BoW; therefore the spatial structures among visual words are ignored. Sivic et al. [6] introduced a new approach based on pLSA for discovering objects and their locations in an image. They introduced 'doublets' which try to consider spatial structural information between visual words. Doublets are two neighboring visual words which have high occurrence probability in each topic. Niu et al. [7] also proposed a new part-based object recognition method based on the supervised LDA in which the spatial information of parts are also considered. It has been shown that BoW models perform poorly on localization [8].

The second group of statistical part-based modeling is called the constellation model. In contrast to BoW model, these constellation models have no independence assumption between parts, and hence exact inference is intractable. In order to simplify the learning and inference procedure, constellation model uses a full multivariate Gaussian distribution to model the spatial distribution of parts [9].

* Corresponding author. Fax: +98 (0) 311 3912718.
E-mail addresses: p.razzaghi@ec.iut.ac.ir, parvin.razzaghi@gmail.com (P. Razzaghi).

1047-3203/$ - see front matter
http://dx.doi.org/10.1016/j.jvcir.2013.12.020

The third statistical part-based modeling is pictorial structures in which the structures between parts are modeled by a tree. In [10], some efficient algorithms for learning these models and matching the model to the image are presented. Methods belonging to this category are divided into three sub-categories based on the method of determining parts. In the first sub-category parts are determined by a human subject which is not useful for most applications [10]. In the second sub-category, parts are determined by a low level interest point detectors. Leibe et al. [11] represented an object as a set of low level features which have a special structure with respect to the center of the object. Their model is called *Implicit Shape Model* (ISM). In [11], at first, low level features, such as Harris corner points [12], are extracted. In the training step, General Hough Transform (GHT) is then used to learn spatial occurrence distribution of the low-level features relative to the object centroid. In the inference step, each feature votes to a point according to the learned spatial occurrence distribution. Consequently, votes are gathered and points whose strengths are greater than a predetermined threshold are selected as an object. Ferrari et al. [13] introduce a new approach for object detection. They use contour feature instead of appearance feature. At first, they learn a prototypical shape of an object class. Then, the learned shape model matches to the object boundaries by using Hough-style voting schema [11,14,15]. The primary weakness of ISM is that each part independently vote whereas there is a mutually dependencies between parts. In [16], this problem has been studied. To do this, in the test phase, groups of features jointly vote to the center of the object. In [16] grouping, voting and correspondences problems are considered jointly and optimized iteratively by a single objective function. Limitation of [16] is that grouping of features is done in the test phase.

In most cases, the used features in ISM do not have enough discriminative information for object classes. Hence, they often match with a background and create false positives. To overcome this difficulty, there are two solutions. Some approaches use a verification stage to overcome this problem. Leibe et al. [11] use an MDL formulation as a verification stage. Yarlagadda et al. [16] apply SVM using pyramid match kernel to verify voting hypothesis. Some other approaches create visual features which have enough discriminative power. Opelt et al. [17] use a Boundary Fragment Model (BFM). They assign a score to each boundary fragment and obtain some of them as candidates. Then, to attain discriminative features, they combined candidate boundary fragments. To do this, boundary fragments should fit well on the positive training samples and their centroid estimates should concur and agree with the true object centroid. It should be noted that, to keep the search tractable, the number of codebook entries should be restricted.

In the third sub-category parts are assumed as latent variables and are achieved during the learning procedure. These parts have a more discriminative ability with respect to the low-level features. Felzenszwalb et al. [18] present an approach in which each deformable object is represented by a filter root and a set of filter parts. In their model, the structure between parts is represented by a spring like graph in which there is a connection between each part and a root filter. Also, for each part, there is a deformation vector which models the deformation of each part with respect to the root filter. In [18] parts are considered as a latent variable and are achieved during the learning procedure. To obtain these latent variables (parts), latent SVM is introduced. Their approach needs an initialization procedure to obtain initial part filters. Also, their approach needs the number of parts to be specified a priori. The structures between parts are also limited to a spring-like graph. Zhu et al. [19] extend [18] to a hierarchical two layer model. Mottaghi [20] integrates the result detection of [18] with Histogram of Oriented Gradient (HOG) bundles [21] to capture large deformations of object. Crandall [1] introduces a new approach in which parts are considered as a latent variable; however the structure between parts is modeled by a *k*-fan graph. Similar to [18], in Crandall's approach an initialization procedure is needed to obtain initial parts appearance. Also the number of parts should be specified a priori. A disadvantage of [1] is that there is no guarantees that model (parts and their relative position) would give a complete description of objects. This is due to the fact that characteristics extracted from the appearance and the structures between the parts are interrelated and are produced concurrently. Hence, the shortcomings in modeling one would affect the other.

In this paper we propose a new part-based modeling approach for object detection. In our approach, for each object category, a set of parts are extracted to comprehensively represent the object and to have strong discriminative ability. Spatial structures of parts of objects are modeled by ISM. In training images, only the bounding boxes of objects are provided and parts are unknown. We propose hierarchical ISM to extract a set of discriminative parts for each object category during the training phase. The proposed hierarchical ISM use a hierarchical star graph to model the structure of each object. In this paper, hierarchical ISM has two layers. In the first layer, each part is modeled by ISM in which the structure of each low-level feature is considered with respect to the part center. A set of ISMs which are used to model parts are called local ISMs. In the second layer, structure of parts with respect to the object center is modeled by a global ISM. Using hierarchical ISM, parts for training images are extracted. To model the visual appearance of each part, a filter on HOG features is learned. In the test phase, each part filter is correlated by test image and candidate positions of all parts are identified. Then, these candidate positions vote to decide on the location of the object center by using global ISM. Each candidate object center is investigated to determine which parts have voted to that position. If an object center position receives votes from all parts, it is considered as a predicted object center. It should be noted that, due to the high discriminative ability of parts, the proposed approach would not require a verification stage.

Our approach differs from others in several important notions: (1) introducing hierarchical ISM to extract a set of discriminative parts for each object category in the training phase, (2) not requiring a verification stage due to high discriminative ability of parts, (3) requiring that all parts vote to a candidate object center location for that candidate to be selected as the object center. This reduces the number of false positive cases. Such requirement is not present in other ISM based methods.

The rest of the paper is organized as follows. The training stage and inference stage of the proposed approach are explained in Sections 2 and 3 respectively. Section 4 shows the results of applying our proposed approach to some known object recognition datasets. Our conclusion is given in Section 5.

2. Proposed approach: training stage

In this section, hierarchical ISM is proposed to extract a set of parts and learn their spatial structure for each object category in the training phase. In training images, only the bounding boxes of objects are provided and parts are unknown. In our approach, at first, a set of initial parts are explicitly determined for each object category using one training image. Then, an initial model for object by using the hierarchical ISM is constructed. Next, based on the initial hierarchical ISM, corresponding of initial parts in other training images are extracted and the hierarchical ISM is updated iteratively. Eventually, to model the visual appearance of extracted parts of training images, a filter is defined using HOG features. The architecture of the proposed approach in the training phase is shown in Fig. 1. Briefly, in the training phase of the proposed approach, the following steps are done: