Neurocomputing 173 (2016) 355–363

Contents lists available at ScienceDirect

Neurocomputing

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

Describing and learning of related parts based on latent structural model in big data

Lei Liu [a], Xiao Bai [b,*], Huigang Zhang [b], Jun Zhou [c], Wenzhong Tang [b]

[a] College of Medical, Shantou University, Shantou 515063 and School of Computer Science and Engineering, Beihang University, Beijing 100191, China
[b] School of Computer Science and Engineering, Beihang University, Beijing 100191, China
[c] School of Information and Communication Technology, Griffith University, Nathan, QLD 4111, Australia

ARTICLE INFO

Article history:
Received 26 June 2014
Received in revised form
25 November 2014
Accepted 14 December 2014
Available online 2 September 2015

Keywords:
Latent structural model
Big data
Incremental learning
Multiple kernel learning

ABSTRACT

In this paper, we propose a novel latent structural model for big data image recognition. It addresses the problem that large amount of labeled training samples are needed in traditional structural models. This method first builds an initial structural model by using only one labeled image. After pooling unlabeled samples into the initial model, an incremental learning process is used to find more candidate parts and to update the model. The appearance features of the parts are described by multiple kernel learning method that assembles more information of the parts, such as color, edge, and texture. Therefore, the proposed model considers not only independent components but also their inherent spatial and appearance relationships. Finally, the updated model is applied to recognition tasks. Experiments show that this method is effective in handling big data problems and has achieved better performance than several state-of-the-art methods.

1. Introduction

Due to the exploding data available on the Internet, tremendous interests in developing big data machine learning methods have emerged in recent years [1–3]. One of the applications in this area is annotation of large scale Web images, for which several datasets containing large number of images collected from the Internet have been created, for instance, Caltech datasets [4], Pascal VOC datasets [5] and ImageNet datasets [6]. Many computer vision and pattern recognition methods have been developed to analyze and annotate big image data. These include classification or ranking methods based on K-Nearest Neighbors (KNN) [7], Support Vector Machines (SVMs) [8], regression models [9], and deep learning [10]. Some methods have used hierarchy strategy to analyze big data problems. Hwang et al learned a tree of semantic kernels, where each node has a Mahalanobis kernel optimized to distinguish the classes in its children nodes [11]. Gao and Koller utilized a hierarchy structure in which image classes are divided into positive and negative groups separated by a binary classifier [12]. In [13,14], the authors also used matrix ideas to perform recognition tasks by dimensional reduction or matrix decomposition.

One of the challenges in dealing with large scale computer vision problems is how to develop effective and efficient feature representation. Many methods adopted the Bag-of-Words (BoW) method [15] which is a vector quantization-based approach. Locally normalized histogram of gradient (HOG) [16] is also a widely used method describe objects, especially for detection tasks [17]. In [18], a fully affine invariant Speeded Up Robust Features (SURF) was proposed to introduce the affine invariant property to SURF feature, while maintaining the feature's own advantage. Recently, Cheng et al. proposed a generic measure for objectness estimation, which is proved to be simple, fast, and effective [19]. Zhao et al. developed a feature coding method based on structural information of local descriptors [20]. In their work, 3D shape context has been incorporated into local spatio-temporal interest point features for human action recognition. Such hierarchical feature coding idea has also been shared by other researchers working on image classification [21].

Besides extracting local features for object description, many methods explore the importance of structural context information in an object, which lead to a batch of structural modeling methods in the literature [22–24]. Lee and Grauman proposed a graph based algorithm that models the interactions between familiar categories and unknown regions, which is used to discover novel categories in unlabeled images [25]. Yang et al. converted an image into a close-loop graph with super pixels as nodes [26]. Saliency can than be determined by ranking these nodes based on their similarities to background and foreground queries. In [27], a multi-

* Corresponding author.
E-mail address: baixiao@buaa.edu.cn (X. Bai).

http://dx.doi.org/10.1016/j.neucom.2014.12.120
0925-2312/

feature fusion method was developed based on semantic similarity for image annotation.

Such structure based methods overcome the shortcomings of missing spatial information in the statistical methods. Part-based model is a special class of structural methods, where structure nodes represent visual parts and graph edges represent the spatial relations between these parts. In [28], Felzenszwalb et al. presented a deformable part model, which combines the part model with latent SVM method to get better recognition results. This method was extended by Ott and Everingham [29], which allows sharing of object part models among multiple mixture components as well as object classes.

Although many part-based methods use human to manually label training samples, in recent years, some efforts have turned to find semantic parts automatically. For example, Singh et al. used an iterative procedure that alternates between clustering and training classifiers, so as to discover a set of discriminative parts [30]. Endres proposed a method to learn a diverse collection of discriminative parts by relocating the object boxes while recognition [31]. Maji and Shakhnarovich presented a method for semi-supervised discovery of semantically meaningful parts from pairwise correspondence annotations [32].

A problem of the part-based methods is that their accuracy may not be guaranteed in case of insufficient labeled training samples. Therefore, some methods used hierarchical or incremental learning schemes to update and enrich the initial trained model [33]. In [34], Zhu and Shao introduced a weakly supervised cross-domain dictionary learning method that uses weakly labeled data from other visual domains as the auxiliary source data for enhancing the initial learning system. Zheng et al. proposed an online incremental learning SVM for large data sets, which consists of learning prototypes and learning support vectors [35]. Chen et al. presented an efficient alternative implementation of incremental learning [36]. It not only improves image processing performance, but also adapts to large datasets. In [37], incremental training of SVM was used as the underlying algorithm to improve the classification time efficiency. Pang et al proposed an incremental learning method that can not only incrementally model the features but also estimate the threshold and training error in a close form [38].

Above all, most of the work done in big data research area have focused on developing fast and efficient recognition algorithms [39]. In this paper, we show how to improve recognition accuracy on top of existing big data techniques. The aim of our work is to develop a latent structured part-based model which uses the inherent relationship between parts to describe objects. Furthermore, our method can extract candidate object parts given only one labeled training image, which is very suitable for big data problems with limited training samples. Different from most structural models, the main contribution of this paper is threefold. Firstly, we propose a novel model formulation that mines the deep relationship (both appearance and spatial relationships) between parts in the object, while most previous works assume parts are independent of each other. Secondly, we present a part finding algorithm which learns a diverse collection of discriminative parts. It only needs one labeled training sample and can save human labelling costs in the training process. Thirdly, we introduce the multiple kernel learning method to describe parts. It enriches the distinctive part information, and therefore, makes the part matching results more accurate.

The rest of the paper is organized as follows. We first present the latent structural model formulation in Section 2. Then we describe the detailed feature extraction and representation method in Section 3. Next, the part finding and learning strategy is introduced in Section 4. Section 5 reports extensive experimental results that validate the effectiveness of the proposed model in big data recognition problems. Finally, we conclude the paper in Section 6 and propose our future work.

2. Latent structural model formulation

In order to predict objects in huge amount of images, we need to build models that can represent these objects. Here we propose a latent structural model that accounts for all the parts and their relationships in the object. This model is enlightened by the discriminative attribute model proposed by Wang and Mori [40] which is a multi-class object classifier that uses attributes as hidden variables. The relationships between the object categories and the attributes are described as learning parameters.

Let a training sample be represented as a tuple (x,h,y), where x is the training image, y is the label, and $h=(h1,h2,\ldots,hm)$ indicates m parts of an object in the image. The classifier $fw:\mathcal{X}\times\mathcal{Y}\rightarrow\mathbb{R}$ is parameterized by vector w. The proposed object model is defined as follows:

$$wT\Phi(x,h,y)=w^{yT}\phi(x;y)+\sum_{j}w^{hj}T\varphi(x;j,hj)+\sum_{(j,k)}w^{jkT}\psi(hj,hk)+\sum_{j}\tilde{w}^{hj}Tv(x;j,hj) \quad (1)$$

In this equation, $w=\{wy,whj,wjk,\tilde{w}hj\}$ are the concatenation of the first parameter in each factor, and the other terms are the features composing $\Phi(x,h,y)$. These terms are defined as follows:

Object classifier $w^{yT}\phi(x;y)$: It is a standard linear model for object recognition without considering object parts. $\phi(x;y)$ is the probability that image x has label y, which can be obtained by training a multi-class SVM.

Part classifiers $w^{hj}T\varphi(x;j,hj)$: It is a standard part model trained to predict the label of part j for image x. It is an independent part for object prediction without considering the object itself or other parts. $\varphi(x;j,hj)$ is the probability that part j is labeled as h_j, achieved by training a binary SVM for this part.

Part/object – part interaction (appearance level) $w^{jkT}\psi(hj,hk)$: It gives the appearance relationship between the j-th part and the k-th part. Furthermore, if we define the 0-th part as the object itself, this model can also give the relationships between the whole object and its parts. $\psi(hj,hk)$ is the joint probability between two parts. It can be achieved by using the part probabilities according to their appearance descriptors.

Object – part interaction (spatial level) $\tilde{w}^{hj}Tv(x;j,hj)$: It represents the spatial relationships between the j-th part and the object center. $v(x;j,hj)$ is a learned variable that gives the location information of part j. In the testing phase, it can be calculated by measuring the spatial overlap ratio between the testing part j and the model part h_j. Furthermore, this term can also describe the spatial relationships between two parts h_k and h_j. Thus, $v(x;j,hj)$ can be re-written as $v(hk,hj)$, which gives the relative locations between each two parts. However, for computational efficiency, we do not use the spatial interactions between parts in this paper.

Although the above model is based on the attribute model in [40], there are some fundamental differences. The goal of our model is describing the part–part relationships and the part–object relationship. First, we propose the part concept instead of attributes, which makes our method more like a latent structural model. Second, we consider the spatial information of parts, and use them to expand the model. This is not a component in [40]. Third, the model in [40] is further modified by combining the object-part and part-part interactions. Last and most important, we develop novel formulation to represent the last two terms in Eq. (1). This allows all four terms use probabilities as measurements. Furthermore, we also introduce an incremental learning process to update the model, which will be discussed later.