Pattern Recognition Letters 50 (2014) 122–129

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Human body part estimation from depth images via spatially-constrained deep learning

Mingyuan Jiu [a,b,*], Christian Wolf [a,b], Graham Taylor [c], Atilla Baskurt [a,b]

[a] Université de Lyon, CNRS, France
[b] INSA-Lyon, LIRIS, UMR5205, Villeurbanne, France
[c] School of Engineering, University of Guelph, Guelph, Ontario, Canada

ARTICLE INFO

Article history:
Available online 14 October 2013

Keywords:
Segmentation
Spatial layout
Deep learning
Convolutional networks
Depth images

ABSTRACT

Object recognition, human pose estimation and scene recognition are applications which are frequently solved through a decomposition into a collection of parts. The resulting local representation has significant advantages, especially in the case of occlusions and when the subject is non-rigid. Detection and recognition require modelling the appearance of the different object parts as well as their spatial layout. This representation has been particularly successful in body part estimation from depth images.

Integrating the spatial layout of parts may require the minimization of complex energy functions. This is prohibitive in most real world applications and therefore often omitted. However, ignoring the spatial layout puts all the burden on the classifier, whose only available information is local appearance. We propose a new method to integrate spatial layout into parts classification without costly pairwise terms during testing. Spatial relationships are exploited in the training algorithm, but not during testing. As with competing methods, the proposed method classifies pixels independently, which makes real-time processing possible. We show that training a classifier with spatial relationships increases generalization performance when compared to classical training minimizing classification error on the training set. We present an application to human body part estimation from depth images.

1. Introduction

Many computer vision problems can be solved in part by an initial step which segments an image, a video, or their constituent objects into regions, which are called parts in this context. The segmentation algorithm typically considers local appearance information, and frequently also models the spatial relationships between different parts. Unfortunately, considering these relationships within the segmentation process mostly amounts to solving constraint satisfaction problems or performing inference in a graphical model with cycles and a non sub-modular energy function, both of which are intractable in the general case. In this paper we address the problem of efficiently modeling spatial relationships without the need for solving complex combinatorial problems.

This general class of problems corresponds to various applications in computer vision. For example, pose estimation methods are also often naturally solved through a decomposition into body parts. A preliminary pixel classification step segments the object into body parts, from which joint positions can be estimated in a second step. The well-known system described in Shotton et al. (2011), installed on millions of gaming consoles and taking as input images from consumer depth cameras, completely ignores spatial relationships between the object parts and puts all of the classification burden on the pixel-wise random forest classifier. To achieve its state-of-the-art level of performance, it required training on an extremely large training set of $2 \cdot 10^9$ examples.

A similar problem occurs in tasks where joint object recognition and segmentation is required. Layout CRFs and their extensions model the object as a collection of local parts (patches or even individual pixels), which are related through an energy function (Winn and Shotton, 2006). However, unlike pictorial structures for object recognition (Felzenszwalb and Huttenlocher, 2005; Felzenszwalb et al., 2010), the energy function here contains cycles which makes minimization more complex, for instance through graph cuts techniques. Furthermore, the large number of labels makes the expansion move-type algorithms inefficient (Kolmogorov and Zabih, 2004).

In all cases, the underlying discrete optimization problem is very similar: an energy function encoding the spatial relationships in pairwise terms needs to be minimized. A typical dependency graph for this kind of problem is shown in Fig. 1b: unary terms relate each label y_i to a feature vector Z_i, and pairwise terms encode

* Corresponding author at: INSA-Lyon, LIRIS, UMR5205, Villeurbanne, France. Tel.: +33 4 72 43 63 72.
E-mail addresses: mingyuan.jiu@liris.cnrs.fr (M. Jiu), christian.wolf@liris.cnrs.fr (C. Wolf), gwtaylor@uoguelph.ca (G. Taylor), atilla.baskurt@liris.cnrs.fr (A. Baskurt).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.09.021

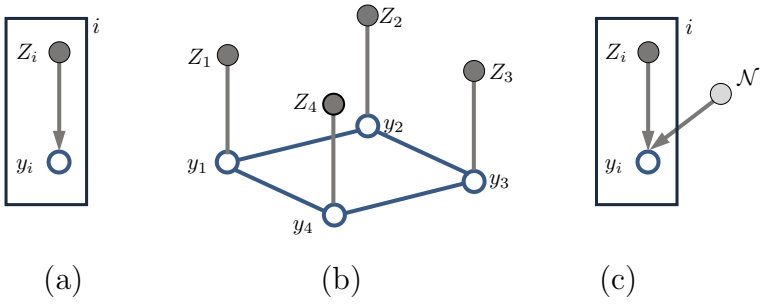

Fig. 1. Different ways to include spatial layout, or not, into learning part labels y_i from features Z_i for pixels i: (a) pixelwise independent classification, where spatial layout information is not taken into account; (b) A Markov random field with pairwise terms coding spatial constraints; (c) our method: pixelwise independent classification including spatial constraints $\mathcal{N}$.

prior knowledge on the possible configurations of neighboring labels y_i and y_j.

In this paper, we propose a method which segments an image or an object into parts through pixelwise classification, integrating the spatial layout of the part labels. Like methods which ignore the spatial layout, it is extremely fast as no additional step needs to be added to pixelwise classification and no energy minimization is necessary during testing. The (slight) additional computational load only concerns learning at an offline stage. The goal is not to compete with methods based on energy minimization, which is impossible through pixelwise classification only. Instead, we aim to improve the performance of pixelwise classification by using all of the available information during learning.

In each of the problems that we consider, the labels we aim to predict have spatial structure. Our proposed method uses an energy function to enforce a spatial consistency in learned features which reflects the spatial layout of labels. Unlike combinatorial methods, our energy function is minimized during training (i.e. while learning features) but is unused at test time. It is based on two main assumptions. First, different high-dimensional features with the the same label are embedded to a lower-dimensional manifold which preserves the original semantic meaning. Second is our belief that greater loss should be incurred when misclassification occurs between features coming from non-neighbor labels than features coming from the same or neighboring labels. In other words, the geometry of learned features, to some extent, reflects the spatial layout of labels. We will show that this new loss function increases the classification performance of the learned prediction model.

Another way of looking at our contribution is to interpret it as a way of structuring the prediction model of a learning machine. Classical techniques working on data represented in a vector space, like neural networks, SVMs, randomized decision trees, boosted classifiers, etc., are, in principle, capable of learning arbitrary complex decision functions if the underlying prediction model (architecture) is complex enough. In reality, the available amount of training data and computational resources available limit the complexity which can be learned. In most cases, only a limited amount of data is available with respect to the complexity of the problem. It is therefore often useful to impose some structure on the model. We already mentioned structured models based on energy minimization and their computational disadvantages. Manifold learning is another technique which assumes that the data, although represented in a high dimensional space, is distributed according to a lower dimensional manifold in that space. Semi-supervised learning uses a large amount of additional training data, which is unlabeled, to help the learning machine better infer the structure of the decision function. In this work, we propose to use prior knowledge in the form of the spatial layout of the labels to add structure to the task of learning the decision function.

Another key aspect of our technique is end-to-end *feature learning*. The dominant methodology in computer vision, though changing in light of recent successes (Krizhevsky et al., 2012), is to extract engineered features such as SIFT (Lowe, 2004) or HOG (Dalal and Triggs, 2005), pool responses, and learn a classifier from this fixed representation. Our objective is to apply learning at all stages of the pipeline, from pixels to labels. However, compared to contemporary Deep Learning approaches, we learn representations that are informed by the spatial structure of the part labels instead of simply their identity.

This paper proposes several contributions:

- We propose spatial learning for unsupervised *pre-training* of deep convolutional networks (i.e. learning all but the topmost layer). The features learned by spatial pre-training are more informative than classical features, as experiments with a non-spatial LR classifier show.
- We propose a framework which integrates spatial part layout into supervised learning of deep neural networks. We show that the gain of spatial learning is indeed obtained at extremely small cost: it improves the performance of the classifier with absolute *zero* increase in computational complexity of testing.
- To the best of our knowledge, we are the first to apply Deep Learning to the problem of body parts segmentation from depth images, obtaining promising results.

2. Related work

Our framework proposes to learn a feature extractor from raw data combining the spatial layout of labels, in order to produce a better decision function for segmentation. It is equivalent to learning a mapping function from high-dimensional space to a low-dimensional manifold space, therefore there is some partial overlap with dimensionality reduction.

Unsupervised approaches for learning a mapping capturing global structure are well-known; most notably Principal Component Analysis (PCA) (Jolliffe, 1986) and Multi-Dimensional Scaling (Cox and Cox., 1994). However, our aim is to embed based on the spatial layout of part labels, so we restrict our discussion to supervised methods. Neighborhood Components Analysis (NCA) (Goldberger et al., 2004) and its variants (Salakhutdinov and Hinton, 2007) implicitly learn a mapping function from high dimensional space to low dimensional space while preserving the neighbourhood relationship defined by class labels. However, NCA is optimized for nearest neighbor classification and does not take into account structure within the labels, only their identity. DrLIM (Hadsell et al., 2006) is an online, non-probabilistic method which explicitly learns a non-linear invariant embedding function. Similarly, we parameterize our embedding with a convolutional neural network (ConvNet) (LeCun et al., 1998), however, like NCA, DrLIM