Signal Processing 110 (2015) 15–26

Contents lists available at ScienceDirect

Signal Processing

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

Human pose estimation via multi-layer composite models ☆

Kun Duan [a],*, Dhruv Batra [b], David J. Crandall [a]

[a] *School of Informatics and Computing, Indiana University, 919 E. Tenth Street, Bloomington, IN 47408, United States*
[b] *Bradley Department of Electrical and Computer Engineering, Virginia Tech, 302 Whittemore, Blacksburg, VA 24061, United States*

ARTICLE INFO

Article history:
Received 5 March 2014
Received in revised form
13 August 2014
Accepted 1 September 2014
Available online 22 September 2014

Keywords:
Object detection
Human pose estimation

ABSTRACT

We introduce a hierarchical part-based approach for human pose estimation in static images. Our model is a multi-layer composite of tree-structured pictorial-structure models, each modeling human pose at a different scale and with a different graphical structure. At the highest level, the submodel acts as a person detector, while at the lowest level, the body is decomposed into a collection of many local parts. Edges between adjacent layers of the composite model encode cross-model constraints. This multi-layer composite model is able to relax the independence assumptions in tree-structured pictorial-structures models (which can create problems like double-counting image evidence), while still permitting efficient inference using dual-decomposition. We propose an optimization procedure for joint learning of the entire composite model. Our approach outperforms the state-of-the-art on four challenging datasets: Parse, UIUC Sport, Leeds Sport Pose and FLIC datasets.

1. Introduction

Detecting humans and identifying body pose are key problems in understanding natural images, since people are the focus of many (if not most) consumer photographs and videos. Accurate pose modeling and recognition could enable a wide range of applications, including detecting suspicious actions in surveillance video, tracking user motion in interactive and immersive video games [24], organizing consumer photo collections automatically based on human activities, and even synthesizing realistic poses in graphics applications [33]. Many current commercial pose estimation systems like Microsoft Kinect use binocular cameras with additional depth sensors, but these techniques do not apply on consumer cameras that do not have these sensors. Detecting people and recognizing pose in 2D static images is significantly more difficult, due not only to the usual complications of object recognition—cluttered backgrounds, scale changes, illumination variations, etc.—but also because of the highly flexible nature of the human body.

Deformable part-based models have emerged as a dominant approach to deal with this flexibility in recognizing people and other articulated objects [10,4,15,35,26,9,31,30]. These models decompose an object into a set of parts, each of which is represented with a local appearance model, and a geometric model that constrains relative configurations of the parts. Recognition is then cast as an inference problem on an undirected graphical model, in which the parts are represented by vertices and the constraints between parts are represented as edges.

Many of these part-based models assume a tree structure, capturing the kinematic constraints between parts of the body—*e.g.* that the lower arm is connected to the upper arm, which is connected to the torso, etc. [10,9,31]. These tree structures allow exact inference to be performed efficiently on the underlying graphical model via dynamic programming. Tree-structured models may seem to be

☆ This is an expanded version of a paper that appeared at the British Machine Vision Conference [7].

* Corresponding author.
E-mail address: kduan@indiana.edu (K. Duan).

http://dx.doi.org/10.1016/j.sigpro.2014.09.014
0165-1684/© 2014 Elsevier B.V. All rights reserved.

ideal for modeling the tree-structured human body, but it turns out that they make assumptions that are not realistic. In particular, the tree structure assumes disconnected parts are conditionally independent, which is not generally true due to constraints imposed by gravity and human balance. For instance, it is much more likely for a person to stand in a symmetrical pose than to stand with all limbs on one side of the torso, but there is no way of encoding this information in a tree model. Similarly, tree models may recognize a single image region as two different body parts, because there is no way to require mutual exclusivity between parts that are not directly connected in the model.

This problem is exacerbated as the number of parts grows: the more parts we have, the more incorrect independence assumptions we are making. (More precisely, note that the number of possible pairs of parts increases quadratically with the number of parts in the model, whereas the number of constraints captured by a tree-structured model increases only linearly.) But recent work suggests that large numbers of parts are beneficial for accurate pose recognition. For instance, Yang and Ramanan [31] found that many small parts with pairwise spatial models based only on translations can approximately model non-affine transformations of the whole object. Thus there is a trade-off between encoding more flexibility in the model by adding parts, and accurately modeling the structural constraints on human pose (Fig. 1).

A variety of approaches have been proposed for overcoming the assumptions made by tree-structured models, including introducing a few cycles into an otherwise tree-structured graphical model [35,30], adding common factor variables [15], or even using a fully connected graphical model [26]. These approaches introduce cycles into the graphical model which generally makes exact inference intractable. How to model richer spatial constraints that still permit efficient inference is an important open question.

In this paper, we propose a new model that addresses these problems from a different perspective. Instead of adding cycles to the original model, we build a multi-level model consisting of multiple tree-structured models with different resolution scales and numbers of parts, allowing different degrees of structural flexibility at different levels, and we connect these models through cross-level links between body parts in adjacent levels. The set of cross-level links also forms a tree. A visualization of our model with three layers is shown in Fig. 2, with cross-level links shown in blue and intra-level links shown in black. The combined graph is no longer a tree, but can be decomposed into tree-structured sub-problems within each level and a cross-model constraint sub-problem across levels. These tree-structured sub-problems are amenable to exact inference, and thus joint inference on the composite model can be performed via dual-decomposition [2]. To learn parameters for our models, we train the cross-layer and intra-layer models jointly, and show that the composite models outperform state-of-the-art techniques on challenging pose recognition datasets. We believe these composite models provide a principled way to trade off between model expressiveness and ease of inference, by "stitching" together multiple tree-structured models into a richer model while keeping the complexity of joint inference in check.

Contributions: To summarize, our contributions include: (1) a novel multi-layer model for human pose estimation, (2) an efficient inference algorithm using dual-decomposition, (3) an algorithm for jointly learning the parameters of our models using structural SVMs, and (4) experimental results showing that the models outperform existing work on modern benchmarks.

Outline: The remainder of this paper presents our method in detail. We discuss related work in Section 2 and describe the model and inference and learning algorithms in Section 3. We evaluate our approach again strong baseline methods in Section 4, before concluding in Section 5.

2. Related work

We begin our summary of related work by describing in Section 2.1 the pictorial structures model, which forms the

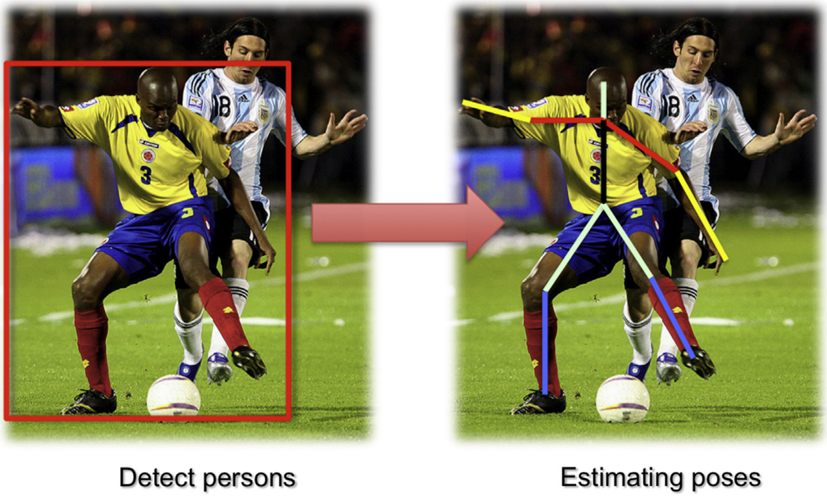

Fig. 1. We consider the problem of detecting humans and estimating their body poses in 2D static images.