Detangling People: Individuating Multiple Close People and Their Body Parts via Region Assembly

Abstract

Today's person detection methods work best when people are in common upright poses and appear reasonably well spaced out in the image. However, in many real images, that's not what people do. People often appear quite close to each other, e.g., with limbs linked or heads touching, and their poses are often not pedestrian-like. We propose an approach to detangle people in multi-person images. We formulate the task as a region assembly problem. Starting from a large set of overlapping regions from body part semantic segmentation and generic object proposals, our optimization approach reassembles those pieces together into multiple person instances. Since optimal region assembly is a challenging combinatorial problem, we present a Lagrangian relaxation method to accelerate the lower bound estimation, thereby enabling a fast branch and bound solution for the global optimum. As output, our method produces a pixel-level map indicating both 1) the body part labels (arm, leg, torso, and head), and 2) which parts belong to which individual person. Our results on challenging datasets show our method is robust to clutter, occlusion, and complex poses. It outperforms a variety of competing methods, including existing detector CRF methods and region CNN approaches. In addition, we demonstrate its impact on a proxemics recognition task, which demands a precise representation of "whose body part is where" in crowded images.

1. Introduction

Person detection has made tremendous progress over the last decade [1]. Standard methods work best on pedestrians: upright people in fairly simple, predictable poses, and with minimal interaction and occlusion between the person instances. Unfortunately, people in real images are not always so well-behaved! Plenty of in-the-wild images contain multiple people close together, perhaps with their limbs intertwined, faces close, bodies partially occluded, and in a variety of poses. A number of computer vision applications demand the ability to parse such natural images into individual people and their respective body parts—for example, fashion [2], consumer photo analysis, predicting interperson interactions [31], or as a stepping stone towards activity recognition, gesture, and pose analysis.

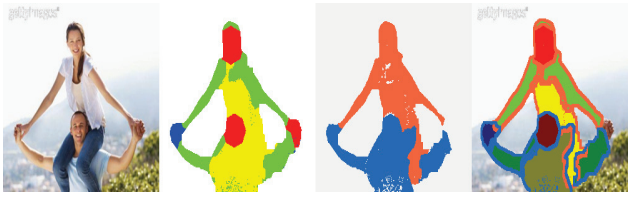

Figure 1. Our method finds human instances and the body part regions (arms, legs, torso, and head). From left to right: input image, semantic body part segmentation, person instance segmentation, final person individuation and part labeling.

Current methods for segmenting person instances [9, 10, 4, 26, 27, 23, 24] take a top-down approach. First they use a holistic person detector to localize each person, and then they perform pixel level segmentation. Limited by the efficiency and performance of person detectors, such methods are slow when dealing with people at unknown scales and orientations. Furthermore, they suffer when presented with close or overlapping people, or people in unusual non-pedestrian-like body poses [31].

We propose a new approach to detangle people and their body parts in multi-person images. Reversing the traditional top-down pipeline, we pose the task as a region assembly problem and develop a bottom-up, purely region-based approach. Given an input image containing an unknown number of people, we first compute a pool of regions using both body-part semantic segmentations and object proposals. Regions in this pool are often fragmented body parts and often overlap. Despite their imperfections, our method automatically selects the best subset and groups them into human instances. To solve this difficult jigsaw puzzle, we formulate an optimization problem in which parts are assigned to people, with constraints preferring small overlap, correct sizes and spatial relationships between body parts, and a low-energy association of body part regions to their person instance. We show that this problem can be solved efficiently using decomposition and a branch and bound