Human and Animal Parsing: Overview

- The goal of human parsing is to detect the positions of the joints and limbs of a human. Thus is needed for many visual tasks. For example, in social interactions the precise configurations of joints/limbs is needed to interpret whether humans are being friendly or aggressive. The precise position of a hand or a foot can make a big change to whether a rockclimber will fall off a mountain or whether a footballer will intercept a football. To know the differences between "man kicks dog" and "dog bites man" requires the ability to parse the human and the dog.
- Moreover human observers have the ability to detect humans, and even parse them into joints/limbs, when only a small fraction of their joints/limbs are visible. This is necessary since, in human interactions, many parts of one human are typically hidden by another human. More abstractly, if a human is seen through apertures (so that large regions of the image are "masked out") then it is still possible for an observer to detect and parse the human, provided many of the joints are visible.
- More generally, we can think of humans (or animals) as being *composed* of elementary parts (joints/limbs) which satisfy spatial relationships. For humans/animals the set of possible configurations is very large and is easiest to specify in three-dimensions (but some regularities apply, e.g., the upper arm is connected to the wrist, the upper legs to the torso, etc.).

Human and Animal Parsing: Deep Networks and Markov Random Fields

- One strategy for representing humans is in terms of graphical models (Markov Random Fields MRFs). Here the nodes of the MRF specify the position of the joints and the pairwise terms (connections between nodes) specify the spatial relations between neighboring keypoints. This is a classic "pictorial structure model". It is formulated using an energy function where the unary terms supply the local evidence for the joints.
- Deep Networks can be used to extend this model. Deep Networks can be used to give evidence for the configurations of the joints as well as their location (e.g., distinguishing between a straight elbow and one bent at ninety degrees). The MRF terms can be used to special spatial and appearance relationships (and can also be learnt). The introduction of these two ingredients (Deep Networks for local configurations, and appearance as well as spatial consistency) gave a huge improvement in the ability to parse humans into joints.
- An alternative way to use Deep Networks to parse humans is to specify architectures (hourglass networks) which output the positions of all the joints, but have no specific representations of them. These methods are trained end-to-end and are very successful on datasets where the amount of occlusion is not too large.

Human and Animal Parsing: Deep Networks and Markov Random Fields

- There are advantages to have models of humans which represent the joints explicitly. In particular, suppose the human is occluded so that only a connected subset of the joints are visible. The human is now represented by many different graphical models (pictorial structures) depending on which subset of joints are visible and we might have to run all of these models independently to parse the human (and to determine which model is right), which is time-consuming. If we represent the joints explicitly, then we can share computation between the different graphical models because we can use the joint-detectors to give evidence for all the different models. Conceptually, we represent the human as a flexible composition of different joints/parts.
- This idea of having multiple graphical models for objects, corresponding to different configurations, with shared components is known as representing the object(s) by compositional models.
- Note: the pictorial structure models (including compositional models) represents the joints in two dimensions. But it is possible to estimate the three-dimensional configuration of the human by using the two-dimensional estimates as input and using prior geometric knowledge about the possible three-dimensional configurations of the humans (beyond the scope of this course).

Human and Animal Parsing: Parsing Animals into Limbs

- We can also represent animals in terms of their limbs (heads, torso, legs, etc.) and the spatial relations between them. These can also be modeled by MRFs where the nodes are limbs, with deep networks trained to give evidence for them, and with the MRF imposing spatial constraints on the relative positions of limbs.
- The limbs can be "shared" between animals. For example, the torsos os horses and cows are fairly similar so we train a cow-horse torso detector which can give evidence for both cow-torsos and horse-torsos. Many limbs/parts can be shared in this manner which reduces the amount of computation and makes the representation more efficient (see later lecture in the course). Some parts, however, such as the head are very distinctive and cannot be shared (and provide cues for distinguishing between different animals).
- These models are also compositional. we can think of an animal are being composed of different parts which can be shared between different objects.
- More complex AND/OR graphs can be used to express different human configurations in terms of a basic vocabulary of elementary parts (this material is in the handout but was not described in the lecture).