Taskonomy: Disentangling Task Transfer Learning

Abstract

Would having surface normals simplify the depth estimation of an image? Do visual tasks have a relationship, or are they unrelated? Common sense suggests that visual tasks are interdependent, implying the existence of structure among tasks. However, a proper model is needed for the structure to be actionable, e.g., to reduce the supervision required by utilizing task relationships. We therefore ask: which tasks transfer to an arbitrary target task, and how well? Or, how do we learn a set of tasks collectively, with less total supervision?

These are some of the questions that can be answered by a computational model of the vision tasks space, as proposed in this paper. We explore the task structure utilizing a sampled dictionary of 2D, 2.5D, 3D, and semantic tasks, and modeling their (1^{st} and higher order) transfer behaviors in a latent space. The product can be viewed as a computational taxonomy and a map of the task space. We study the consequences of this structure, e.g., the emerging task relationships, and exploit them to reduce supervision demand. For instance, we show that the total number of labeled datapoints needed to solve a set of 10 tasks can be reduced to $\frac{1}{4}$ while keeping performance nearly the same by using features from multiple proxy tasks. Users can employ a provided Binary Integer Programming solver that leverages the taxonomy to find efficient supervision policies for their own use cases.

1. Introduction

We intuitively understand that surface normals and depth are related, and that the walls of a room mark vanishing points that are useful for orientation. Other task relationships are less clear: such as how point matching is related to curvature or understanding the mechanism through which keypoint detection and the lighting in a room can, together, do pose estimation.

Although adapting *input* domains and modeling their relationships is recognized as an important concept and is often employed (domain adaptation), modeling the relationships between tasks, i.e. *output* spaces, is harder and consequently often ignored.

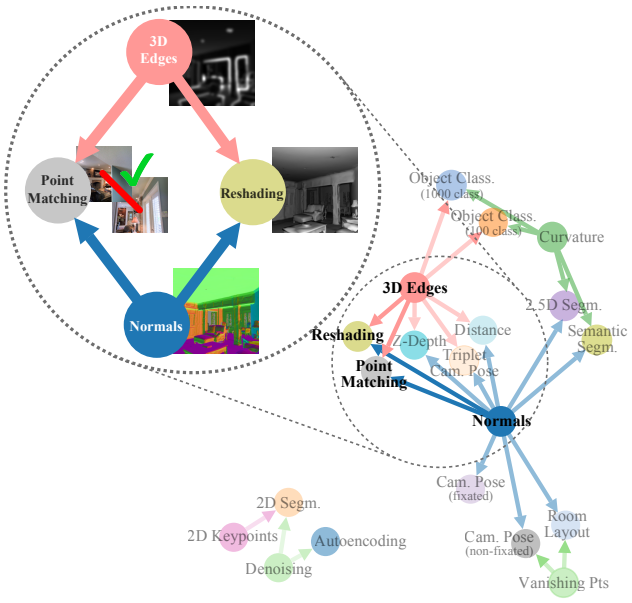

Figure 1: **A sample task structure discovered by the computational taxonomy is shown.** For example, it discovers that by combining the learned features of a surface normal predictor and occlusion edge detector, good networks for reshading and point matching can be rapidly trained.

The field of computer vision has indeed gone far without explicitly using these relationships. We have made incredible progress by developing advanced learning machinery (e.g., CNNs) that is capable of finding complex mappings from A to B when the pairs (a,b) are exposed. This is referred to as **fully** supervised learning, and this formulation often leads to problems being solved in isolation.

This siloing makes training a general-purpose system into a Sisyphean challenge, whereby each new task needs to be learned from the very beginning. It ignores these quantifiably useful relationships means that a fully supervised approach requires enormous amounts of data. Alternatively, a model which uses the structure in the task space needs less supervision, may be more robust, uses less computation, and behaves in more predictable ways. This structure and its effects are still largely unknown. The relationships are non-trivial, and finding them is complicated by the fact that the we have imperfect models and optimizers. In this paper we attempt to shed light on this underlying structure. We