Scene Collaging: Analysis and Synthesis of Natural Images with Semantic Layers

Phillip Isola
MIT
phillipi@mit.edu

Ce Liu
Microsoft Research
celiu@microsoft.com

Parsing an image into a set of objects and interactions remains a grand challenge in computer vision. When humans look at a scene, e.g., a cityscape, a forest, or a cafeteria, we see an organized interaction of objects, functions, and spaces. However, many existing scene parsing approaches represent an image simply as a 2D array of pixel labels (e.g. [6], [5], [2], [7]), and these representations fail to account for occlusion. In typical images, huge swathes of scene structure are occluded from view. Further, occlusion even makes visible content difficult to parse: when projected into a 2D image, background objects are fragmented by occluders.

To solve these problems, we have proposed a novel scene model, in which we represent discrete semantic objects on separate layers (published in full length as [4]). We take our lead from human artists. While there are many ways by which an artist can synthesize a scene, one of the quickest and easiest is to collage together the image out of found pieces. Following this collaging approach, we model a scene as a collage of object segments sampled from a large database of example images for which humans have provided object label annotations (Figure 1).

These “scene collages” are a type of layer-based model. Layer models have been widely used in both computer vision and graphics: e.g., for video representation [9], image synthesis [1], and segmentation [10]. In the context of scene parsing, however, only very recent and concurrent work has explored layer-based representations [3], [8].

Our full pipeline is shown in Figure 2. Formally, a scene collage is a generative model of an image. Given a query image, we infer the collage that best explains it through analysis-by-synthesis. First, from a dataset of labeled exemplar scenes, we retrieve a set of candidate object segments that visually match the query image. Then, we combine elements of this set into a collage to fully explain the query. Each object segment in the collage is placed so as to match the visual appearance of some region of the query.

Contextual coherence between the object segments is enforced through the use of a scene prior. This prior operates on the “scene graph” of a scene collage. The scene graph is a structure whose nodes represent which objects are in the scene and whose edges represent physical support relationships. Our scene prior is a grammar over valid scene graphs. The intuition behind the grammar is that we interpolate between the scenegraphs of exemplar scenes in our dictionary. We consider the following set of interpolation moves: \{birth, death, swap\}. Each move takes a given valid scenegraph and restructures some part of it to be more like the scenegraph of an exemplar scene (Figure 3). Intuitively, a birth move adds an object to our scenegraph, a death move removes an object, and a swap exchanges one object for another. Any scene whose scenegraph can be generated in this way is considered to be valid according to our grammar. All other scenegraphs – such as one in which a “mountain” is supported by a “sun” – are considered invalid and given a prior probability of zero. Inference proceeds greedily: starting with an empty collage, at each step we select the valid move that most improves the visual match between the collage and query image.

We apply our system to both traditional scene parsing
The primary contribution of our work is the application of a novel layer-world representation to understanding and manipulating natural images. Our scene collage representation naturally consists of geometric relationships such as occlusion and depth ordering, and could be further extended to a full 3D model. For more details, please see our full length paper [4].