This CVPR2015 extended abstract is the Open Access version, provided by the Computer Vision Foundation.

Part-based modelling of compound scenes from images

Anton van den Hengel[1], Chris Russell[2], John Bastian[1], Daniel Pooley[1], Anthony Dick[1], Lachlan Fleming[1], Lourdes Agapito[2]
[1] The Australian Centre for Visual Technologies,The University of Adelaide
[2] University College London

We propose a method to recover the structure of a compound scene from multiple silhouettes. Structure is expressed as a collection of 3D primitives chosen from a pre-defined library, each with an associated pose. This has several advantages over a volume or mesh representation both for estimation and the utility of the recovered model. The main challenge in recovering such a model is the combinatorial number of possible arrangements of parts. We address this issue by exploiting the intrinsic structure and sparsity of the problem, and show that our method scales to scenes constructed from large libraries of parts.

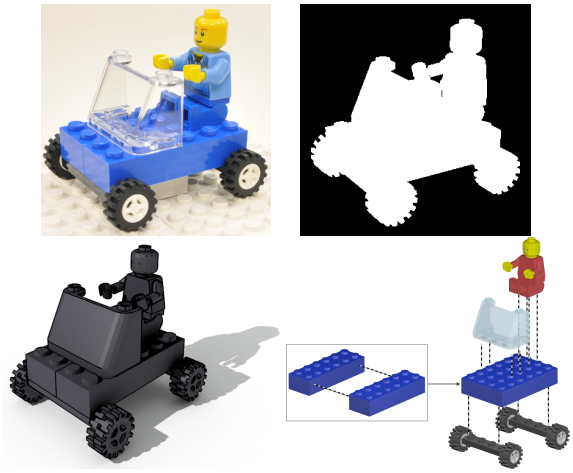

Figure 1: An illustration of the modelling process, from real world images, and silhouettes, to an estimate of the building blocks from which an object is constructed, and how they fit together.

Unlike most work on the recovery of shape from images, our method does not generate a point cloud, or a volume, but a structural explanation of way the scene depicted is constructed. In this sense it is aligned with the blocks-world approach [2], recently revisited by [1]. Much like the blocks-world approaches, our goal is to recover a semantic model of the structure of the scene. Instead of creating a simpler volumetric, or point cloud model of a scene, we wish to create a model which captures interdependencies between parts of a scene, and allows us to say "These are the wheels of the car so this is how it will move." (fig. 1), or "This is how a wall might collapse in an accident, or a temple might collapse in an earthquake".

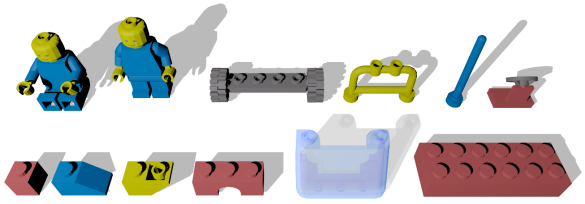

Figure 2: A selection of the template types used in recovering the structure of the Lego® models. The set of templates consists of each template type rendered in every plausible location and orientation.

The method we propose reasons in 3D about the structure of a scene on the basis of its appearance in an image set. This requires an initial set of building blocks from which an scene might be composed. As we are interested in structure, rather than appearance, these building blocks are defined uniquely by their shape and position. Our recovered structure estimate is the smallest set of building blocks required to reconstruct the scene in question.

We formulate the problem of physically plausible structure from silhouette as a mixed integer programme. Soft constraints that the discovered structure must share the same silhouette are formulated as linear constraints as are hard constraints that blocks must be placed in a physically plausible structure that is self-supporting with no selected blocks intersecting one other. Finally, hard binary constraints that eliminate fractional solutions and make sure that a block is either fully selected or completely ignored are imposed to tighten the linear programme and guarantee a physically plausible solution.

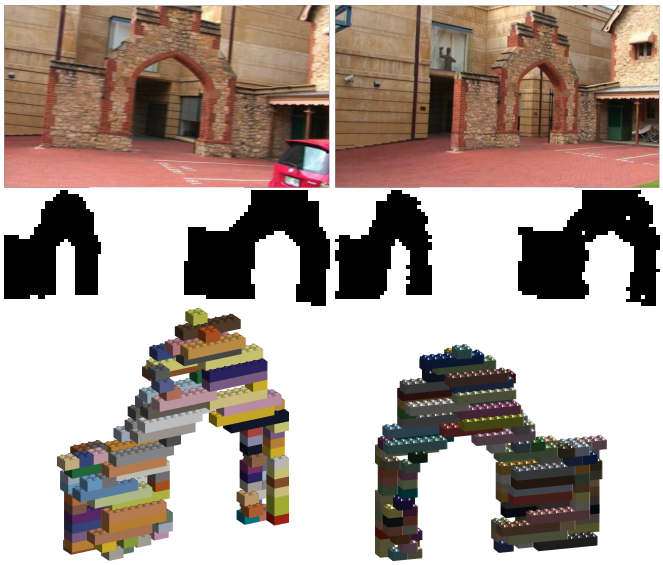

Figure 3: Two input images of the Archway sequence their silhouettes, and those of the reconstructed model, followed by two rendered images of the model.

Figure 4: Two views of a model constructed from 'The Dinosaur Sequence' of Wolfgang Niem, University of Hannover.

[1] Abhinav Gupta, Alexei A. Efros, and Martial Hebert. Blocks world revisited: Image understanding using qualitative geometry and mechanics. In *ECCV*, 2010.
[2] L. Roberts. *Machine perception of 3D solids*. PhD thesis, Stanford, 1965.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.