Int J Comput Vis (2013) 104:38–68
DOI 10.1007/s11263-013-0613-4

Recovering Relative Depth from Low-Level Features Without Explicit T-junction Detection and Interpretation

Felipe Calderero · Vicent Caselles

Received: 9 August 2012 / Accepted: 10 January 2013 / Published online: 7 February 2013

Abstract This work presents a novel computational model for relative depth order estimation from a single image based on low-level local features that encode perceptual depth cues such as convexity/concavity, inclusion, and T-junctions in a quantitative manner, considering information at different scales. These multi-scale features are based on a measure of how likely is a pixel to belong simultaneously to different objects (interpreted as connected components of level sets) and, hence, to be occluded in some of them, providing a hint on the local depth order relationships. They are directly computed on the discrete image data in an efficient manner, without requiring the detection and interpretation of edges or junctions. Its behavior is clarified and illustrated for some simple images. Then the recovery of the relative depth order on the image is achieved by global integration of these local features applying a non-linear diffusion filtering of bilateral type. The validity of the proposed features and the integration approach is demonstrated by experiments on real images and comparison with state-of-the-art monocular depth estimation techniques.

Keywords Low-level image features · Monocular depth · Relative depth order · Multi-scale analysis

1 Introduction

The human visual system is able to perceive depth even from a single and still image thanks to monocular (static) depth cues (Goldstein 2002). These cues are also known as pictorial cues, since they have been used by painters for centuries in their search for realistic effects in their masterpieces.

In the context of computer vision, the monocular depth estimation problem can be defined as inferring the depth order of the objects present in a scene using only information from a single view or image. Except in cases where the size of the objects is known a priori, in general, the extracted depth relations are relative, meaning that it can only be concluded that an object is closer or farther (or at the same depth) to the camera than another object, but no information about its absolute depth position can be extracted.

Monocular depth cues have been extensively studied by psychologist during the last century, using psycho-visual and perceptual studies (Howard 2012). These cues can be mainly classified in three types: high-level, mid-level, and low-level cues. High-level cues do involve image understanding from a semantic point of view, and usually require previous object extraction and recognition stages to be inferred. Some of the most relevant are relative and known size, or light and shadow distribution.

We refer to mid-level cues as those that require a basic knowledge of the image content. Some examples are perspective, that requires detecting first the presence of projective distortions of the scene; texture gradient, based on previous extraction of textured areas of the image; or atmospheric perspective, that can be only applied in landscapes or large depth range images.

On the contrary, low-level cues do not necessarily require a high level analysis of the image (although they can benefit from it) and can be directly and locally inferred. Among them, let us mention blur, convexity, closure, and intersection (also known as occlusion) cues. Blur appears in objects outside the depth of field of the camera (the distance between the nearest and farthest objects in a scene that appear acceptably sharp).

F. Calderero (✉) · V. Caselles
Universitat Pompeu Fabra, Roc Boronat, 138, 08018 Barcelona, Spain
e-mail: felipe.calderero@upf.edu

V. Caselles
e-mail: vicent.caselles@upf.edu