Image and Vision Computing 27 (2009) 514–522

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Local detection of occlusion boundaries in video

Andrew N. Stein *,[1], Martial Hebert

The Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA

ARTICLE INFO

Article history:
Received 22 January 2007
Received in revised form 16 January 2008
Accepted 24 April 2008

Keywords:
Occlusion boundaries
Edge detection
Contours
Motion analysis
Occlusion detection

ABSTRACT

Occlusion boundaries are notoriously difficult for many patch-based computer vision algorithms, but they also provide potentially useful information about scene structure and shape. Using short video clips, we present a novel method for scoring the degree to which occlusion is visible at detected edges. We first utilise a spatio-temporal edge detector which estimates edge strength, orientation, and normal motion. By then extracting patches from either side of each detected (possibly moving) edge pixel, we can estimate and compare motion to determine if occlusion is present. In experiments on synthetic and natural images, we demonstrate our ability to differentiate occlusion boundary pixels from simple edge pixels by using motion information. In terms of precision versus recall, our occlusion scoring metric outperforms a rank-based motion inconsistency measure from the literature. The completely local, bottom-up approach described here is intended to provide powerful low-level information for use by higher-level reasoning methods.

1. Introduction

Occlusion boundaries provide strong cues about the 3D structure of a natural scene. Their detection has application in scene segmentation, figure-ground separation, and shape extraction, all of which can improve object recognition and detection (*e.g.* [1]). Such boundaries correspond to locations in an image where one physical surface is closer to the camera than another and are usually also visible as appearance edges. In this paper, we explicitly distinguish between the detection of typical *edges*, which may result from changes in intensity, colour, or texture, and *boundaries* which additionally correspond to 3D scene structure. Indeed, we will consider occlusion boundaries to be a subset of the edges in a scene, as demonstrated in Fig. 1. The goal of this work is to identify those edges in an image which are also occlusion boundaries. As with general scene segmentation, measuring the performance of edge detectors is fairly ill-defined beyond the use of semantic labels provided by humans. Occlusion boundaries, however, have a physical meaning for which the notion of a "correct" answer is more easily defined.

Occlusion cannot be directly observed from a single image. Thus, while the use of sophisticated edge detectors that can handle texture and colour [2–4] may offer improved results over a simple Canny edge detector [5], they still (correctly) respond strongly to edges which do not correspond to any physical occlusion. Occlusion can, however, be observed through motion – either the scene's, the camera's, or both. Without motion (and without non-trivial higher-level knowledge), it is impossible to distinguish between edges and boundaries based solely on *local* information in a *single* image.

When one object occludes another, due to either object's motion or to parallax from camera movement, pixels may disappear or become visible. This occlusion and disocclusion are the source of significant difficulty for many computer vision methods, which often rely on image patches that may overlap the boundaries. Since it is generally assumed that all the pixels within a patch "belong together" (*e.g.* are from the same object, motion layer, *etc.*), patches overlapping occlusion boundaries violate this assumption and muddy results. For this reason, patches near boundaries are often treated as outliers, or multiple/adaptive-windowing techniques are employed [6–8]. By contrast, the work of this paper will focus precisely on these boundaries themselves and will attempt to detect them directly by augmenting standard single-image edge detection with motion information.

We avoid dense motion estimation and subsequent region-growing or clustering by focusing explicitly on motion *at the edges* in a scene. In addition, we can detect *local* occlusions directly using a bottom-up approach which is complementary to top-down approaches that rely on higher-level reasoning. Such methods often impose the restrictive assumption that the scene consists of a set of distinct layers moving separately (and often that the number of layers is known). Alternatively, the challenge of a purely local, bottom-up approach is estimating and utilising motion information at exactly those locations (occlusion boundaries) at which it is arguably the most difficult to obtain.

* Corresponding author. Tel.: +1 412 268 1420.
E-mail addresses: anstein@cmu.edu (A.N. Stein), hebert@ri.cmu.edu (M. Hebert).

[1] Partial support provided by a National Science Foundation Graduate Research Fellowship, with additional support from the Intelligent Robotics Development Program, a 21st Century Frontier R&D Program funded by the Korean Ministry of Commerce, Industry, and Energy.

0262-8856/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2008.04.017

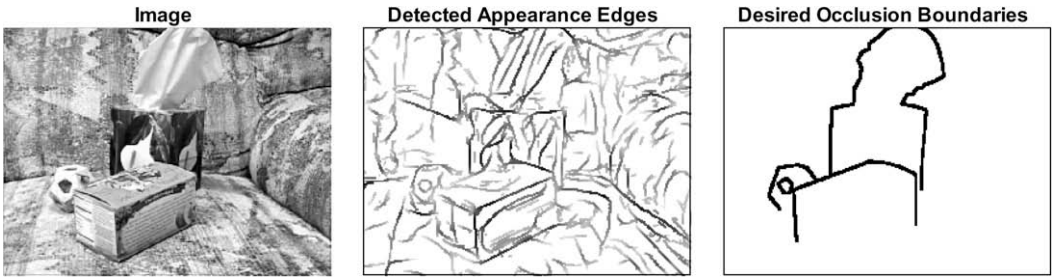

Fig. 1. For the scene on the left, an appearance-based edge detector may produce the result shown in the middle. But only a subset of the detected edges correspond to occlusion boundaries (shown at right), which carry additional 3D shape and structure information that is likely more useful for higher-level reasoning about the scene.

Using only a few frames of video (typically 6–10), we will employ space-time edge detection [9–12] in Section 4, which simultaneously estimates local edge strength, orientation, and crucially, edge *speed* in the direction normal to its orientation. Given these quantities, we can then safely extract spatio-temporal patches of intensity data from either side of a moving edge. By analyzing and comparing the motion in these two patches, we can estimate the degree to which we believe occlusion is occurring there, as described in Section 5 (and following an earlier version of this work [13]). We emphasise that our approach involves small relative motions visible within a few frames (*e.g.* due to parallax from small camera motions) and that it requires no explicit tracking of pixels/features over time.

2. Related work

Martin et al. [2] have designed an excellent edge detector which has been trained (using human-labeled data) to respond to those local gradients of brightness, colour, or texture which people generally seem to label as edges. (Note that the authors' use of the term "boundaries" in that work does not directly correspond to the extraction of the occlusion boundaries sought in this work.) Their comparison of histograms on either side of proposed oriented edges allows the detection of difficult complex edges, such as those between textured regions or in clutter. The Compass Edge Detector uses a very similar histogram-based approach [3,4] but without learning all the parameters from human-labeled data. More recently, similar ideas using non-parametric tests of distributions within a patch were explored in [14] for detecting texture edges. None of these approaches, however, incorporate motion information or seek specifically to identify occlusion boundaries.

Smith et al. [15] track edge fragments and use Expectation Maximization to then segment the scene into regions with consistent motion. Their approach is one of the few that tracks edges/boundaries directly, but it still assumes layered scene structure, and the method seems to work best on two-layer sequences (computation increases exponentially with the number of layers). Our completely local approach, on the other hand, makes no assumption that the scene is layered and should therefore be applicable to a more general set of scenes. Also, the initial step of linking edge pixels into chains, which are then assumed to be part of a single surface, is often considered an "implementation detail" for many methods requiring boundary-fragments, including [15]. But this task is non-trivial and quite crucial since it imposes a hard decision on the remainder of the system.

Black, Fleet, and Nestares [16–18] also attempt to estimate local evidence of occlusion by analyzing motion. They build a parametric model of local occlusion within sampled circular regions and then estimate the posterior probability of this model using particle filtering. Demonstrated results are fairly limited, possibly due to the significant computational expense of evaluating the thousands of particles necessary to sample the parameter space for each region.

Many approaches segment dense motion estimates derived from optical flow or feature tracking into distinct regions or layers (*e.g.* [19–22], to name just a few), usually treating the erratic results at boundaries as outliers to an underlying smooth process. The subsequent delineation of precise motion boundaries, if performed at all, is generally of secondary importance. A notable exception, however, is found in [23], where vertical and horizontal between-pixel motion boundaries plus their interactions with nearby dense optical flow vectors are considered in a Markov Random Field (MRF) framework. Stereo or structure from motion techniques also have trouble near occlusion boundaries and usually focus on the interiors of regions while handling data near occlusions as complex special cases [6–8]. Our work, on the other hand, is not concerned with precise dense motion estimation or full 3D scene reconstruction; we seek only to *identify* oriented boundary locations that correspond to visible occlusion. Such information should be useful for higher-level reasoning, *e.g.* by a feature-based object recognition method which utilises boundary knowledge [1].

Browstow and Essa [19] segment a static scene into planar relative depth layers by observing the occlusions introduced by an object moving through that scene. The use of edge detection and contour completion help provide good estimates of the layers' (occluding) boundaries. The approach, which is limited to a stationary camera, is aimed primarily at compositing video and may also require manual intervention.

Also related to occlusion detection is the classical problem of T-junction detection, recently addressed in a discriminative framework by [24] (see references therein for substantial prior work). Building on the Epipolar Plane Image idea [25,26], their work utilises spatio-temporal slices (not volumes) and demonstrates limited extraction of occlusion boundaries. For higher-level reasoning, note that T-junction detection may be a complementary source of information to the motion *boundary* detection presented here.

As discussed above, our approach will compare motions from two patches of spatio-temporal data to determine their consistency. This exact problem was recently addressed, though in a rather different context, in the work of [27]. That approach utilises a continuous rank-increase measure between Gram matrices constructed from spatio-temporal derivatives within each patch. This measure, to which we will compare our approach in the results section, provides a motion inconsistency score without explicitly estimating the patches' motion vectors.

3. Properties of local occlusion

Assume for a moment that we have detected a small edge fragment in an image. We expect the patches on either side of a