Pattern Recognition 64 (2017) 15–28

Contents lists available at ScienceDirect

Pattern Recognition

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

Joint occlusion boundary detection and figure/ground assignment by extracting common-fate fragments in a back-projection scheme

Cheng Chen[a], Jason J. Corso[b,*]

[a] Oklahoma State University, United States
[b] Electrical Engineering and Computer Science, University of Michigan, 1301 Beal Avenue, Ann Arbor, MI 48109, United States

ABSTRACT

Occlusion boundary detection and figure/ground assignment are among the fundamental challenges for the real world visual pattern recognition applications, such as 3D spatial understanding, robotic navigation and object search. We attack these challenges by extracting an intermediate-level image/video representation, namely, Common-Fate Fragments. A Common-Fate Fragment is composed of both over-segmented region and edge fragments. Physically, it exists as a coupled edge-region fragment bound with dynamic information. Common-Fate Fragment candidates are generated by an integrated line-region growing process, which does not require complete object segmentation or closed object boundary extraction. To identify Common-Fate Fragments from these extracted candidates, we introduce a back-projection verification scheme that can circumvent the notoriously difficult task of direct motion estimation on boundaries. This allows occlusion detection and figure/ground labeling to be jointly conducted within a simple but effective hypothesize-and-test framework. We test the proposed method on YouTube Motion Boundaries (YMB) data set and two benchmark data sets: the CMU and Berkeley motion data sets. Even though the idea of the proposed method is simple and transparent, promising experimental results are observed.

1. Introduction

Each frame of a video is a perspective projection of the 3D world, which is full of opaque objects occupying different depths. Objects that are spatially closer to a camera will occlude, either entirely or partially, the objects that are further away. In typical videos of the real-world, the occurrence of occlusion is the norm rather than an exception [38]. Therefore, detection of occlusion due to depth discontinuity is crucial for pattern recognition and computer vision to operate in the real world.

Occlusion boundaries and appearance edges are two distinct conceptions. Following the established definition [17], an appearance edge refers to the typical output of an edge detection algorithm on intensity or color image data, whereas an occlusion boundary is explicitly created by objects covering one another. After decades of development, the performance of appearance edge detection algorithms is becoming closer and closer to human performance on benchmark datasets [4,42]. Occlusion boundaries typically occur at appearance edges, but a detected appearance edge is in no way sufficient to guarantee an occlusion boundary [17]. Once an appearance edge has been formed, it needs to be identified as a boundary or not. If it is a boundary, then we may ask which side of it is the figure (object) and which side is the occluded background.

The human vision system can easily deal with this kind of boundary detection and figure/ground labeling tasks. Psychophysical research shows that the visual system makes use of occlusive relations in the real world to recover depth, contour, and surface [13]. A Gestalt principle, “common-fate” states that objects moving together should be grouped together. Evidence suggests that this simple strategy is unconsciously used for occlusion reasoning and figure/ground assignment in human vision [24]. According to this rule, if an edge moves together only with a region on one of its sides, it should belong to that side region (the figure/foreground region) [7,25]. If the regions on both sides of a given edge move together, most likely this edge is an appearance edge instead of an occlusion boundary.

However, without an accuracy motion estimation on an edge, it is not a trivial task to judge whether a given edge moves together with only one of the regions on its sides. Brightness constancy and spatial smoothness are two common assumptions that underlie typical optical flow estimation methods [10]. The state-of-the art optical flow estimation methods already can achieve satisfactory results when these assumptions are met. However, they encounter problems at occlusion

[*] Corresponding author.
E-mail address: jjcorso@eecs.umich.edu (J.J. Corso).

http://dx.doi.org/10.1016/j.patcog.2016.10.013
Received 15 June 2016; Received in revised form 23 September 2016; Accepted 10 October 2016
Available online 20 October 2016

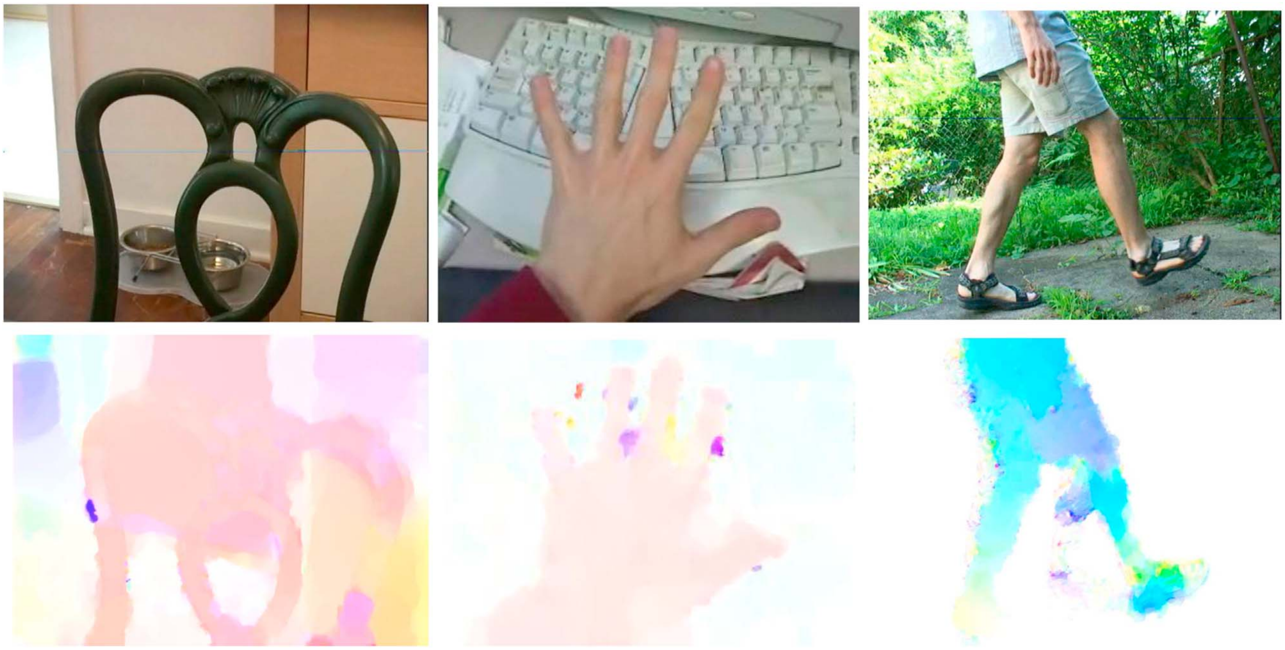

Fig. 1. The first row shows input video sequences from a bench mark CMU data set [33]. The second row shows the visualized optical flows with color-coding [8]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

boundaries where the assumption of spatial smoothness is violated. As a result, motion estimates are often imprecise and blurry in these areas as demonstrated by Fig. 1. Even though many efforts [18,36] have been made to prevent blurring of optical flow across image boundaries, reliable motion estimation at object boundary places is still notoriously difficult [8,34].

To tackle the above challenges, we introduce an intermediate edge-region representation and back-projection verification scheme, which can circumvent the direct motion estimation for boundary areas. The intermediate representation is called Common-Fate Fragments and explicitly groups edges with regions on either side, which we refer to as *side-regions*. These edge-region pairs facilitate the reasoning about occlusion detection and figure/ground labeling, which is conducted in a simple but effective hypothesize-and-test framework.

Inspired by the part-whole hierarchy perceptual organization principle [27], we link our intermediate Common-Fate Fragment representation directly to the back-projection verification scheme. Psychologists believe that raw visual signals are first grouped into some elemental visual units, which are composed of both over-segmented regions and edges [26,27], then structured features are extracted for building more semantically meaningful visual entities. We believe that complete object segmentation and closed object boundary extraction are not necessary and also not achievable for a low or middle-level computer vision process because “what you see is what you need.” In other words, these ill-defined problems can only be solved in a high-level process guided by specific hypotheses. Therefore, without the need of complete object segmentation or closed object boundary extraction, we extract both region and edge fragments (building blocks of Common-Fate Fragment candidates) simultaneously through an integrated line-region growing process. The extracted region fragments facilitate a filtering process to eliminate these motion outliers for a reliable affine motion model estimation.

Based upon an estimated affine motion model on one side-region fragment, the edge fragment can be back-projected to the previous frame of the video stream. If the edge fragment is moving together with the side-region fragment, the back-projected edge fragment will match the edge map on the previous frame very well. By comparing the fitness of two back projections (using the affine motion models estimated from both side-region fragments), we can infer with which side-region fragment that an edge fragment is moving. Therefore, we can apply the “common-fate” rule to identify an occlusion boundary and assign figure/ground labels to its two attached regions simultaneously.

2. Related work

Occlusion boundary detection and figure/ground assignment are long standing research topics in pattern recognition and computer vision. There are numerous related literatures. Here, we only review the most directly related work which explicitly exploits motion cues.

Prior attempts to use motion cues for occlusion boundary detection can be traced back to the early work of motion layers [2]. The key idea is to fit each moving object into a motion layer so that motion segmentation can be obtained by assigning each pixel into different layers. Some researchers, such as, Wang and Adelson [40], and Bergen and Meyer [6], identify such occasions by checking whether a group of pixels on the edge of a motion layer are outliers. Most of these motion layer methods need a global multi-layer parametric motion model. However, the model parameter estimation is a difficult task without knowing the model order. It has been noted that these methods are sensitive to the accuracy of optical flow and that accurate optical flow is hard to estimate without prior knowledge of the occlusion boundaries [34]. Therefore, occlusion detection and optical flow estimation become “chicken and egg problems.” Unknown motion layers and noisy optical flow estimation around occlusion boundary regions are the two difficulties faced by these motion layer methods. To attack these challenges, Ayvaci et al. [5] formulate occlusion detection and optical flow estimation as a joint optimization problem.

Based on the assumption that occlusion boundaries occur at the static edges, the more recently developed methods extract static edge cues as candidates for occlusion boundary detection. Moreover, without the assumption of the number of motion layers or moving objects, local motion models have been verified to work well in handling motion and occlusion for both challenging synthetic and real video streams [34]. In order to generate candidate edges and extract local motion features, optical flow estimation and over-segmentation are commonly conducted as a pre-process [15,29,33,34]. The performance of these methods largely depends on two factors: proper edge candidate selection and reliable local motion model estimation. In general, these