Computer Vision and Image Understanding 133 (2015) 66–75

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

# Focus-aided scene segmentation ☆

Said Pertuz [a,*], Miguel Angel Garcia [c], Domenec Puig [b]

[a] Escuela de Ingeniería Eléctrica, Electrónica y Telecomunicaciones, Universidad Industrial de Santander, Bucaramanga, Colombia
[b] Department of Computer Science and Mathematics, Universitat Rovira i Virgili, Tarragona, Spain
[c] Department of Electronic and Communications Technology, Autonomous University of Madrid, Madrid, Spain

ARTICLE INFO

*Article history:*
Received 11 June 2014
Accepted 30 September 2014
Available online 13 October 2014

*Keywords:*
Image sequences
Focus measure
Segmentation
Defocus
Visual cue integration
Focus cue

ABSTRACT

Classical image segmentation techniques in computer vision exploit visual cues such as image edges, lines, color and texture. Due to the complexity of real scenarios, the main challenge is achieving meaningful segmentation of the imaged scene since real objects have substantial discontinuities in these visual cues. In this paper, a new focus-based perceptual cue is introduced: the *focus signal*. The focus signal captures the variations of the focus level of every image pixel as a function of time and is directly related to the geometry of the scene. In a practical application, a sequence of images corresponding to an autofocus sequence is processed in order to infer geometric information of the imaged scene using the focus signal. This information is integrated with the segmentation obtained using classical cues, such as color and texture, in order to yield an improved scene segmentation. Experiments have been performed using different off-the-shelf cameras including a webcam, a compact digital photography camera and a surveillance camera. Obtained results using Dice's similarity coefficient and the pixel labeling error show that a significant improvement in the final segmentation can be achieved by incorporating the information obtained from the focus signal in the segmentation process.

## 1. Introduction

The visual system is fundamental for humans in order to perceive the surrounding environment. At the lowest level of scene understanding, the task of distinguishing different objects is fundamental due to its key role in the interaction with the perceived world. In particular, segmenting a scene into different objects can be understood as finding the spatial relationship between them and their distance or depth from the observer. In his seminal work on physiological optics, von Helmholtz distinguishes two main sources of visual information or *visual cues* [1]: the first source relies on experience and some familiarity with the nature of the perceived scene. Some mechanisms corresponding to this source are the determination of distance by means of the relative size of objects, their perspective, texture patterns and shading. The second source involves and actual perception of depth, such as *vergence*, depth perception by motion parallax (e.g., by moving the head), *stereopsis* (binocular or stereo vision), and accommodation (or focusing).

Albeit each visual cue can be regarded as an independent source of information, at a higher complexity level, object segmentation is a complex task that involves the combination and interaction of different cues. As a result, in order to yield a deeper understanding of the human visual system, it is critical not only to understand each individual cue, its principles, limitations and advantages, but also to understand their integration and interaction.

The focus cue in both human and computer vision is an increasingly important research field. Interestingly enough, the literature concerning the integration of the focus cue with classical perceptual cues, such as color, texture, shading, etc., is scarce. In this work, a framework for the efficient integration of the focus cue with other low-level monocular cues in order to yield an improved scene segmentation is proposed. Specifically, this paper presents a new interpretation for the autofocusing process by exploiting the implicit information about the scene geometry found in the variations of the focus level yielding and improved image segmentation. The claim is that, being autofocus an unavoidable part of systems with limited depth-of-field, it is possible to take advantage of this process in order to infer useful information about the imaged scene.

Based on an analysis of the image formation process of a defocused image, we propose to model autofocus as a time-variant interaction between the capturing device and the observed scene, showing that each imaged point generates a pattern, or *focus signal*,

☆ This paper has been recommended for acceptance by Y. Aloimonos.

* Corresponding author.

*E-mail addresses:* spertuz@uis.edu.co (S. Pertuz), miguelangel.garcia@uam.es (M.A. Garcia), domenec.puig@urv.cat (D. Puig).

http://dx.doi.org/10.1016/j.cviu.2014.09.009
1077-3142/© 2014 Elsevier Inc. All rights reserved.

that mainly depends on the configuration of the lens-camera system and the scene geometry. Since at every instant, the lens-camera system has the same configuration for all imaged points, the scene geometry (in particular the separation of objects in space) can be estimated as a function of the different *focus signals*, where the focus signals correspond to the focus measure as a function of time.

For illustration, Fig. 1 shows an image stack of a video sequence recorded while a camera is autofocusing on a real scene. The scene is divided into a discrete number of local regions of interest. An initial focus-based segmentation of the scene is recovered by clustering the focus signals extracted from each region of interest. The *signal clusters* depicted in Fig. 1 can be interpreted as a segmentation process through which the image is segmented into disjoint regions by taking into account the geometry of the scene rather than the color or texture features typically used in segmentation tasks. Finally, the segmentation is refined by incorporating texture and color cues by means of classical segmentation schemes.

This paper is organized as follows. Section 2 reviews previous related work. Fundamental concepts for the proposed approach, namely *focus measure* and the *focus signal*, are introduced in Sections 3 and 4. The methodology for exploiting the focus signal in order to yield an improved scene segmentation is presented in Section 5. The proposed approach is experimentally evaluated in Section 6. Section 7 is a final discussion.

## 2. Previous work

Since the work by von Helmholtz [1], the role of accommodation and defocus in visual perception has extensively been assessed in the literature of human vision. More recently, it has been experimentally shown that, in fact, the observed amount of defocus blur is an independent human pictorial depth cue by itself [2,3]. These results suggest that, in addition to classical cues (stereopsis, shading, texture, perspective, etc.), the perceived blur is exploited for retrieving information about the structure of the scene, such as the depth of objects and occlusions.

In computer vision, different researchers have tackled the problem of integrating different visual cues. Early efforts to integrate different low-level visual cues can be traced back to [4], where the joint effect of stereo, shading and texture was analyzed in depth perception. In this scope, the existing approaches have mostly dealt with the integration of low- and high-level perceptual cues not related to focus, such as contours, optical flow and image features [5,6]. More recently, Sabatini et al. integrated optical flow, contrast orientation and binocular disparity for depth perception [7]. In [8], Ogale and Aloimonos proposed a framework for the cooperative integration of disparity with basic visual models, such as segmentation, shape and depth estimation, and occlusion detection. Some researchers have devoted their efforts on assessing the integration of visual cues, such as perspective and disparity [9,10] or texture and shading [11] for specific tasks, such as detecting surface orientation. Similarly, Pang et al. integrated disparity, color and shape for real-time object tracking [12]. Recent approaches are aimed at the integration of multi-sensor data for performing basic visual tasks such as visual tracking and simultaneous location and mapping [13,14]. In cooperative cue integration for improved perception, perceptual grouping and stereo disparity have also been studied [15].

Most research involving the focus cue has been devoted to depth estimation through *shape-from-focus* (SFF) and *shape-from-defocus* (SFD). SFF consists in capturing an ordered sequence of images of the same scene with different camera settings in order to change the degree of focus of every imaged scene point. A depth-map can be estimated by measuring the focus level of each point of the scene whereas the position at which each point presents maximum focus will determine its position with respect to the camera [16,17]. In SFF, it is assumed that the lens-camera configuration (in terms of the in-focus position) is known for each captured scene. SFF has been applied in microelectronics, fracture analysis, polymeric texture analysis and the reconstruction of microscopic objects [18–22].

SFD is based on the principle that, in diffraction limited optics, depth can be estimated by measuring the amount of defocus [23–25]. SFD aims at determining depth as a function of the relative defocus between observations. There have been different approaches in both the frequency domain [26] and the spatial domain of images [25,27] for estimating depth from two or more images. More recently, Zhuo and Zim have proposed a method to obtain a defocus map from a single image [28]. SFD, as well as SFF, perform poorly in scenes with low texture content and have often been applied in controlled environments.

If the camera's hardware is allowed to be modified and manipulated, different focus-based alternatives that have provided good results in focus-based depth estimation are the coded aperture [29,30] and the plenoptic imaging [31] approaches. With an rigurous control of the camera's optics and configuration (namely, the lens focal length, aperture and exposure time), light-efficient photography has been also studied [32]. From the very beginnings of the shape-from-defocus framework, Engelhardt and Knop exploited active illumination in order to perform real-time focus-based depth estimation [33]. Subsequently, active depth recovery by projection of light patterns has been exploited in [34] and, more recently, in [35].

The literature specifically regarding the combination of focus with other perceptual cues is relatively scarce. In particular, [36] proposed a hybrid system that combines stereo and shape-from-focus in order to estimate depth. To the best of our knowledge, the problem of integrating the focus cue with classical visual cues such as texture and color for scene segmentation has not been tackled previously.

## 3. Defocus and focus measure

In order to obtain a deeper understanding of the focus cue, it is necessary to analyze how focus and defocus are perceived by

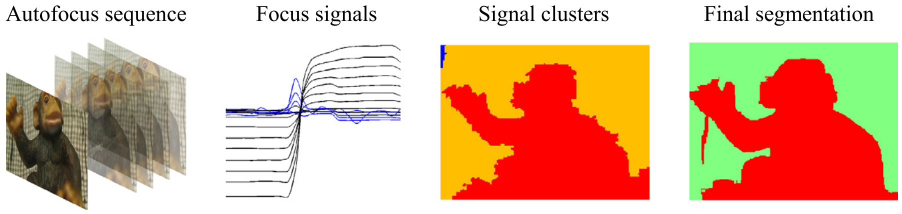
Autofocus sequence
Focus signals
Signal clusters
Final segmentation

**Fig. 1.** Focus-aided scene segmentation. An initial focus-based segmentation is generated by clustering the focus signals extracted from the image frames of an autofocus sequence (signal clusters). The focus signals correspond to the focus measure of non-overlapping image regions as a function of time. The initial focus-based segmentation is then refined by incorporating classical segmentation cues, such as texture and color, in order to yield an improved focus-aided scene segmentation (final segmentation). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)