J. Vis. Commun. Image R. 39 (2016) 1–11

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Depth extraction of partially occluded objects using deformable net ☆

Khaled M. Shaaban [a,*], Nagwa M. Omar [b]

[a] Electrical Engineering Department, Faculty of Engineering, Assiut University, Assiut, Egypt
[b] Information Technology Department, Faculty of Computers and Information, Assiut University, Assiut, Egypt

ARTICLE INFO

Article history:
Received 30 July 2015
Revised 7 January 2016
Accepted 3 May 2016
Available online 7 May 2016

Keywords:
Monocular vision navigation
Depth extraction
Partially occluded objects
Deformable net
Video segmentation

ABSTRACT

Estimating depth using a single camera moving along its optical axis has great benefits to autonomous robot navigation. In Shaaban and Omar (2012) we estimated depth information for non-occluded regions using the change of their seen area. The proposed work extends this technique to handle regions with partial occlusion. As the robot moves, the seen area of a partially occluded region is different from the two points of view. Therefore, at least one of the edges is deceptive and is not a real object boundary. We propose a technique to detect deceptive-edges at the boundary of occlusion then use triangulation to detect the structure of occlusion. With this knowledge, we estimate the percentage change in the actual area seen as the camera moves. This percentage is used to correct the results of the original Area-based method. Experimentally the algorithm provides an average depth with good accuracy for both far and near objects.

1. Introduction

Estimating depth information is an essential task in computer vision. It has important applications in scene understanding and robotic navigation. The most popular technique for estimating depth information is binocular vision (stereopsis) [2–13]. However computing high-quality dense depths is still a challenging problem particularly if we consider the influence of image noise, textureless regions, and occlusions that are inherent in the image capturing process [13,14]. More importantly, the accuracy declines with the ratio between the baseline length and the distance to the object.

Other techniques are proposed to estimate depth using a single image or a sequence of images acquired under a relative motion of a single camera, which is called Monocular Vision [1,9,13–32]. Without any prior information, estimating the depth of a scene from a single image is a highly ambiguous problem. Due to the ambiguities inherited in the problem, many of the existing methods rely on supervised learning; this makes the outcomes heavily depend on the training data sets [9,15,18,19,24,31]. Some of the proposed methods that estimate the depth from single still image model depth only at a local scale and still lack reasoning about the global structure of the scene [15–17,26]. In [27] the image is analyzed at coarse scale to infer the scene structure but it fails to provide an absolute depth estimate. Similar to [26], the work in [27] ignores the task of semantic understanding which puts an enormous burden on the learning algorithm. The authors in [24] try to overcome this disadvantage by proposing an approach that reasons about the semantic content of a scene. However, the technique achieves poor performance on the objects that are not in the training set.

Obviously video sequence has more information about the scene. Therefore, some of the techniques rely upon video sequences captured by rotating a single camera around the object or by arbitrary camera/object movement such as camera pan or zoom as shown in [18,19,22]. The proposed methods in these works converted conventional monocular videos to stereoscopic ones which required days and sometimes months for conversion. For example in [19], the authors performed automatic foreground segmentation and determine the relative depth changes of moving objects by analyzing the optical flow between frames. The system combines the prior analysis through motion and the depth ordering by user interaction to generate dense depth maps. The authors in [18] tried to infer the optimal depth by minimizing the energy defined on a Markov Random Field. The system adopts a trained model to predict pixel disparities based on their motion and depth. The authors in [22] also proposed a method that converts 2D videos into stereoscopic 3D by combining motion and photometric cues together to estimate depth map.

Our point of view is that if the motion is perpendicular to the optical axis, then mathematically we could consider the techniques that use video sequences captured by rotating a single camera around the object as stereo vision but with the two images are not acquired concurrently. System with such motion constrain is

☆ This paper has been recommended for acceptance by M.T. Sun.

* Corresponding author.
E-mail address: k_shaaban@hotmail.com (K.M. Shaaban).

http://dx.doi.org/10.1016/j.jvcir.2016.05.004
1047-3203/

with limited use to robotic navigation. Also the techniques that coverts 2D video to 3D video in days and months are suitable for Cinema production not for real time applications such as robot navigation. A more useful and difficult case is the one where the motion is parallel to the optical axis which we proposed in [1]. However, the analysis in this case is difficult. It depends upon partitioning the first frame into its segments [33] then tracking these segments across the subsequent frames. The change of the segment areas as the camera moves forwarded provides the means to estimate the distance to the corresponding objects [1]. This technique gives good results since it uses the whole area of the object, which makes the estimation of the distance immunes to noise and less sensitive to camera resolution. Furthermore, it is particularly useful in the case where the distance to the object is in hundreds of meters, which is the case where stereo vision gives poor results. This concept is the one we follow in this work. Unfortunately, this technique is very sensitive to occlusion. For example, consider the case where a robot carrying a single camera is moving parallel to the optical axis with a partially occluded object in scene. In this case, the actual area seen from the object as the robot moves forward changes with distance, therefore, depending upon the change of the projection of this area in the image leads to wrong results. Compensating for occlusion when using the area based distance estimation is the main goal of this work. Therefore, we will review some of the main techniques to handle occlusion in image analysis.

In case of stereo vision, the occlusion is taken into consideration while estimating a disparity map [4–7,10,11]. Occlusion detection in stereo vision is performed implicitly [10,11] or explicitly [5,7,12]. In implicit detection, occlusions are identified during the matching process. In the explicit detection, occlusions are first discovered by left to right and right to left matching. Then the two-way disparity values of corresponding points are compared [5,7]. Inequality of magnitudes of these disparity values indicates occlusion. In a consequent step the disparity of occluded points are estimated during occlusion filling algorithm [2]. Most of these algorithms are proven to be computationally expensive [2,4–7,10–12].

Also, there are many research work tried to handle the occlusion problem in case of Monocular vision [9,16,20,24,26,27,30–32]. In [9], Saxena and others introduced method to estimate depth based on integrating monocular and stereo cues to model depths and relationships between depths at the points in the image for occluded and non-occluded objects using Markov Random Field. At small distances, the algorithm depends more on stereo cues while at larger distances the algorithm relies more on monocular cues from one image. The monocular cues rely on prior knowledge learned from a training set. Thus, the monocular cues may not be useful for images outside the training set. Saxena and others in [26] extend the work in [9] to estimate 3-D structure from a single still image of an unstructured environment using a Markov Random Field (MRF) that is trained via supervised learning. This technique computes features to predict significant boundaries in the images, such as occlusion and folds. The technique encodes some higher-level information by modeling the relationships of neighboring superpixels, but it requires more work on the global structure of the scene. Liu and others [24] proposed technique motivated by the work of Saxena and colleagues [26,31] for 3-D reconstruction that makes use of semantic information to guide depth perception. The authors avoid the need for modeling occlusions and they obtain them from the semantic labels.

Some of the research work extends the depth extraction from single image to videos taking into consideration the occlusion problem. The authors in [16] used local motion cues to improve the inferred depth maps, while optical flow is used to ensure temporal depth consistency. The technique models the relationships of neighboring superpixels, but it requires analysis about the scene global structure. Raza and others [17] propose learning a mapping of monocular cues to depth using statistical learning and inference techniques. They infer the depth information of new scenes using piecewise planar parameterization estimated within a Markov random field (MRF) framework. They combined appearance, motion, occlusion boundaries, and geometric context of the scene to predict depth. Stein and Hebert [25] also has shown that combining appearance and motion cues improves occlusion boundary detection. They further improve occlusion boundary detection by applying a global conditional random field. Recently, Sundberg et al. [23] also tried to improve occlusion boundary detection by computing motion gradients across static boundaries. Since the methods in [25,23] rely on local features they are unable to reduce false positives where intra-object local motion or appearance variance is high.

In [1] we introduced structure from motion technique to autonomously extract objects' 3D information from video sequence captured from a single camera moving parallel to its optical axis without any prior knowledge, training or user interaction. We utilized the RbDN strategy that we developed in [33] to achieve real time Video Segmentation that deforms an elastic-net that represents the contours of the different segments of the images. This net has the form of a planer graph that carries the information of all segments including corners, average colors, neighboring segments and more. The technique in [1] deforms this elastic net to track the changes in the segments through the frames. The correspondence of the segments and their corners across the frames is automatically tracked which eliminates the need for solving the correspondence problem. Avoiding the correspondence problem provides the speed needed for real time applications such as robotic navigation. From the corresponding position of the vertices, 3D information about the objects in the scene could be obtained using method we called Vertex-Based. Using the ordinary triangulation, sensitivity of the estimated 3D information for the points near the optical axis is shown to be small which leads to low accuracy. To overcome this problem we proposed another method called Area-Based to get the average distance of the different regions in the scene. This method depends upon the changes in the segment area with the motion of the camera to estimate the object distance. The Area-Based method is mathematically proven more accurate and experimentally provided improved results in the cases with no occlusion. In the case of partially occluded objects, the object area that could be seen from the camera changes with the change of its point of view. This change should be considered when using the area change to estimate the distance to the object.

In this paper, we propose a technique that extends the work in [1] to estimate the average depth to partiality occluded and non-occluded objects autonomously in real time without user interaction, training, prior knowledge or assumption about the contents of the sequence. The first step is to locate real edges that represents real object boundary. Then we track the change in the position of these edges in two frames to estimate the extra area covered or revealed due to the change in the point of view of the camera. This extra area is used to correct the results of the Area-Based method.

The problem we are trying to solve is more useful in robotic application than the techniques used to convert the monocular video to stereoscopic ones in weeks [18,19,22]. In addition, unlike the techniques used to estimate depth from single image [9,15,18,19,24,31], the proposed technique works autonomously without user interaction, training or prior knowledge. Also, compared with stereo vision techniques [2–8,10–13], the proposed technique could be used to estimate the depth to both close and faraway objects. The technique eliminates the need to solve the correspondence problem, which allows for real time performance needed for autonomous robot navigation.