International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163
Issue 2, Volume 2 (February 2015) www.ijirae.com

Depth Extraction from Video: A Survey

Akshay Chavan
PG Scholar, Dept of ECE
Reva Institute of Technology and Management
Bangalore, India

Raveendra S Gudodagi
Assistant Professor, Department of ECE
Reva Institute of Technology and Management
Bangalore, India

Abstract--- The proposed work is the literature survey done on Depth extraction from a video using many approaches. A detail study is made on the various algorithms used and develops the techniques that automatically generate plausible depth maps from the videos. The technique which is applicable to both single images as well as video is presenteds. The survey also deals with the techniques that can be used to automatically convert the monoscopic video into the stereo for 3D visualization. An algorithm for depth estimation from a monocular video sequence containing moving and deformable objects is also presented

Keyword--- Bilayer segmentation, depth recovery, motion estimation, image-based modeling/rendering, molecular vision, bundle optimization.

I. INTRODUCTION

Scene depth is useful for a variety of tasks, ranging from 3D modeling and visualization to robot navigation. It also facilitates spatial reasoning about objects in the scene, in the context of scene understanding. In the growing 3D movie industry, knowing the scene depth greatly simplifies the process of converting 2D movies to their stereoscopic counterparts.

While many reconstruction techniques for extracting depth from video sequences exist, they typically assume static scenes and moving cameras. They do not work for dynamic scenes or for rotating, stationary or variable focal length sequences. There are some exceptions, which can handle some moving objects, but they still require camera motion to induce parallax and allow depth estimation.

In this paper, we present novel solutions to generate depth maps from ordinary 2D videos; our solution also applies to single images. These techniques are also applicable to arbitrary videos, and works in cases where conventional depth recovery methods fail (static/rotating camera; change in focal length; dynamic scenes).

II. LITERATURE SURVEY

A. Robust Bilayer Segmentation and Motion/Depth Estimation with a Handheld Camera

The method of extracting high-quality dynamic foreground layers from a video sequence is a difficult problem due to the coupling of color, motion, and occlusion [1]. Many approaches assume that the background scene is static or undergoes the planar perspective transformation. These restrictions are reduced and a comprehensive system for accurately computing object motion, layer, and depth information is developed. A novel algorithm that combines different clues to extract the foreground layer is proposed, where a voting-like scheme robust to outliers is employed in optimization. The system is capable of handling challenging examples in which the background is nonplanar and the camera freely moves during video capturing. Many applications such as high-quality view interpolation and video editing is achieved.

The commonness of Internet video facilitates the development of video editing techniques. Layer extraction is one of the essential tools that allow users to separate foreground and background images in a video. Its application lies in the ability to produce layers whereby users can easily create special effects, such as inserting the foreground object into a virtual environment, and to accomplish necessary adjustment, including removing unwanted objects or deforming them.

High-quality layer separation from a video is a very challenging problem because tightly coupled color, depth, and motion give rise to a large number of variables and significant inexactness in computation. Previous methods made various assumptions on the background scene or camera motion to simplify the problem. For example, bilayer segmentation methods assume that the camera is fixed and/or the background color distribution is not complex. It is, however, very common that a captured video does not meet these requirements. So, methods that can relax these conditions are in demand.

Background modeling is a vital step for bilayer segmentation. If the background image is known, the foreground estimate can be obtained with the color and contrast information .Otherwise, both layers have uncertain pixel assignments, making accurately identifying them challenging. The latter scenario commonly arises when using a handheld camera.

In this approach, solution for the layer segmentation problem with the input of only a video sequence taken by a freely moving camera is presented. The objective is the high-quality dynamic foreground extraction, which requires that the computed layers have accurate and temporally consistent boundary in multiple frames. In addition, dense motion fields and

depth maps need to be solved for. To accomplish these goals, this method uses several new measures and contributes an iterative optimization scheme to refine the depth and motion estimates.

In the bilayer segmentation step, depth and motion are used to handle layer occlusion and resolve color similarity. Unlike traditional solutions that weight different terms in an objective function, we employ a simple voting-like strategy to effectively balance the set of terms and automatically reject occasional outliers. The bilayer segmentation result is used to refine the optical flow field on the dynamic foreground, avoiding the errors caused by connecting a foreground pixel with a background one.

B. Tour into the Picture

TIP (Tour into the Picture) is used for making animations from one 2D picture or photograph of a scene [2]. In TIP, animation is created from the viewpoint of a camera which can be three-dimensionally "walked or flownthrough" the 2D picture or photograph. To make such animation, conventional computer vision techniques cannot be applied in the 3D modeling process for the scene, using only a single 2D image. Instead a spidery mesh is employed in this method to obtain a simple scene model from the 2D image of the scene using a graphical user interface. Animation is thus easily generated without the need of multiple 2D images.

Unlike existing methods, this method is not intended to construct a precise 3D scene model. The scene model is rather simple, and not fully 3D-structured. The modeling process starts by specifying the vanishing point in the 2D image. The background in the scene model then consists of at most five rectangles, whereas hierarchical polygons are used as a model for each foreground object. Furthermore a virtual camera is moved around the 3D scene model, with the viewing angle being freely controlled. This process is easily and effectively performed using the spidery mesh interface.

Doing animation from one picture, painting, or photograph is not a new idea. Such animations have been mainly used for techniques are fully available. When the input image is given in advance, first of all, the animator has to make the 3D scene model by trial and error until the projected image of the model fits well with the input image of the scene.

In this approach a simple method, which we call TIP (Tour into the Picture), for making animations from one 2D picture or photograph of a scene. This method provides a simple scene model, which is extracted from the animator's mind. Thus the scene model is not exactly 3D structured, but is geometrically just a collection of "billboards" and several 3D polygons. Therefore, the animations obtained with this method are not strictly three-dimensional. However, the proposed method allows easy creation of various animations, such as "walk-through" or "fly-through", while visually giving convincing 3D quality.

C. Image-Based Modeling and Photo Editing

In this approach an image-based modeling and editing system that takes a single photo as input [3]. This method represents a scene as a layered collection of depth images, where each pixel encodes both color and depth. Starting from an input image, this approach employs a suite of user-assisted techniques, based on a painting metaphor, to assign depths and extract layers. Here two specific editing operations are introduced. The first, a "clone brushing tool," permits the distortion-free copying of parts of a picture, by using a parameterization optimization technique. The second, a "texture-illuminance decoupling filter," discounts the effect of illumination on uniformly textured areas, by decoupling large- and small-scale features via bilateral filtering. This system enables editing from different viewpoints, extracting and grouping of image-based objects, and modifying the shape, color, and illumination of these objects.

Despite recent advances in photogrammetry and 3D scanning technology, creating photorealistic 3D models remains a tedious and time consuming task. Many real-world objects, such as trees or people, have complex shapes that cannot easily be described by the polygonal representations commonly used in computer graphics. Image-based representations, which use photographs as a starting point, are becoming increasingly popular because they allow users to explore objects and scenes captured from the real world. While considerable attention has been devoted to using photographs to build 3D models, or to rendering new views from photographs, little work has been done to address the problem of manipulating or modifying these representations. This method describes an interactive modeling and editing system that uses an image-based representation for the entire 3D authoring process. It takes a single photograph as input, provides tools to extract layers and assign depths, and facilitates various editing operations, such as painting, copypasting, and relighting.

In this approach, The extension for photo editing to 3D is done. This method describes a system for interactively editing an image-based scene represented as a layered collection of depth images, where a pixel encodes both color and depth. This system provides the means to change scene structure, appearance, and illumination via a simple collection of editing operations, which overcome a number of limitations of 2D photo editing.

D. Automatic Photo Pop-up

This approach presents a fully automatic method for creating a 3D model from a single photograph [4]. The model is made up of several texture-mapped planar billboards and has the complexity of a typical children's pop-up book illustration. The main insight is that instead of attempting to recover precise geometry, here statistically model geometric classes are defined by their orientations in the scene. This algorithm labels regions of the input image into coarse categories: "ground", "sky", and "vertical". These labels are then used to "cut and fold" the image into a pop-up model using a set of simple assumptions. Because of the inherent ambiguity of the problem and the statistical nature of the approach, the algorithm is not expected to work on every image. However, it performs surprisingly well for a wide range of scenes taken from a typical person's photo album.

In this approach, a method for creating virtual walkthroughs that is completely automatic and requires only a single photograph as input. This approach is similar to the creation of a pop-up illustration in a children's book: the image is laid on the ground plane and then the regions that are deemed to be vertical are automatically "popped up" onto vertical planes. Just like the paper pop-ups, the resulting 3D model is quite basic, missing many details. Nonetheless, a large number of the resulting walkthroughs look surprisingly realistic and provide a fun "browsing experience".

The target application scenario is that the photos would be processed as they are downloaded from the camera into the computer and the users would be able to browse them using a 3D viewer and pick the ones they like. Just like automatic photo-stitching, this algorithm is not expected to work well on every image. Some results would be incorrect, while others might simply be boring. This fits the pattern of modern digital photography – people take lots of pictures but then only keep a few "good ones". The important thing is that the user needs only to decide whether to keep the image or not.

E. A dynamic Bayesian network model for autonomous 3D reconstruction from a Single indoor image

When we look at a picture, the prior knowledge about the world allows us to resolve some of the ambiguities that are inherent to monocular vision, and thereby infer 3D information about the scene. We also recognize different objects, decide on their orientations, and identify how they are connected to their environment. Focusing on the problem of autonomous 3D reconstruction of indoor scenes, in this approach we come across a dynamic Bayesian network model capable of resolving some of these ambiguities and recovering 3D information for many images[5]. This model assumes a "floorwall" geometry on the scene and is trained to recognize the floor-wall boundary in each column of the image. When the image is produced under perspective geometry, This model can be used for 3D reconstruction from a single image. To our knowledge, this was the first monocular approach to automatically recover 3D reconstructions from single indoor images.

Given only a single image, depth estimation cannot be done by geometry-only approaches such as a straightforward implementation of stereopsis. In this approach, the focus is made exclusively on 3D reconstruction from a single indoor image. Motivation for studying this problem is two-fold. First, to anticipate that monocular vision cues could later be applied in conjunction with binocular ones; however, restricting our attention to monocular 3D reconstruction allows us to more clearly elucidate what sorts of monocular vision cues are useful for depth estimation. The second motivation is that monocular vision to be interesting and important in its own right. Specifically, monocular cameras are cheaper, and their installation is less complex than, stereo cameras. More importantly, the accuracy of stereo vision is fundamentally limited by the baseline distance between the two cameras, and performs poorly when used to estimate depths at ranges that are very large relative to the baseline distance. Straightforward implementations of stereo vision also tend to fail in scenes that contain little texture, such as many indoor scenes (that contain featureless walls/floors). In these settings, monocular vision may be used to complement, or perhaps even replace standard stereopsis.

F. Learning Depth from Single Monocular Images

In this approach, the task of depth estimation from a single monocular image is considered [6]. A supervised learning approach to this problem is taken into consideration, in which we begin by collecting a training set of monocular images (of unstructured outdoor environments which include forests, trees, buildings, etc.) and their corresponding ground-truth depthmaps. Then, supervised learning to predict the depthmap as a function of the image is applied. Depth estimation is a challenging problem, since local features alone are insufficient to estimate depth at a point, and one needs to consider the global context of the image. This model uses a discriminatively-trained Markov Random Field (MRF) that incorporates multiscale local- and global-image features, and models both depths at individual points as well as the relation between depths at different points. This approach show that, even on unstructured scenes, this algorithm is frequently able to recover fairly accurate depthmaps.

Recovering 3D depth from images is a basic problem in computer vision, and has important applications in robotics, scene understanding and 3D reconstruction. Most work on visual 3D reconstruction has focused on binocular vision (stereopsis) and on other algorithms that require multiple images, such as structure from motion and depth from defocus. Depth estimation from a single monocular image is a difficult task, and requires taking into account the global structure of the image, as well as using prior knowledge about the scene. In this approach, supervised learning to the problem of estimating depth from single monocular images of unstructured outdoor environments, ones that contain forests, trees, buildings, people, buses, bushes, etc is applied.

This approach is based on capturing depths and relationships between depths using an MRF. The method began by using a 3D distance scanner to collect training data, which comprised a large set of images and their corresponding ground-truth depthmaps. Using this training set, the MRF is discriminatively trained to predict depth; thus, rather than modeling the joint distribution of image features and depths, this method model only the posterior distribution of the depths given the image features.

G. Make3D: Learning 3D Scene Structure from a Single Still Image

In this approach, the problem of estimating detailed 3D structure from a single still image of an unstructured environment is considered [7]. The main goal is to create 3D models that are both quantitatively accurate as well as visually pleasing. For each small homogeneous patch in the image, a Markov Random Field (MRF) is used to infer a set of "plane parameters" that capture both the 3D location and 3D orientation of the patch. The MRF, trained via supervised learning, models both image depth cues as well as the relationships between different parts of the image. Other than assuming that the environment is made up of a number of small planes, our model makes no explicit assumptions about the structure of the scene; this enables the algorithm to capture much more detailed 3D structure than does prior art and also give a much richer experience in the 3D flythroughs created using image-based rendering, even for scenes with significant non-vertical structure. Using this approach, qualitatively correct 3D models for 64.9 percent of 588 images are achieved which is downloaded from the Internet.

In this approach, the main goal is to infer 3D models that are both quantitatively accurate as well as visually pleasing. Most 3D scenes can be segmented into many small, approximately planar surfaces. (Indeed, modern computer graphics using OpenGL or DirectX models extremely complex scenes this way, using triangular facets to model even very complex shapes.) This algorithm begins by taking an image and attempting to segment it into many such small planar surfaces. Using a superpixel segmentation algorithm, an oversegmentation of the image that divides it into many small regions (superpixels) is obtained.

H. Repetition-based Dense Single-View Reconstruction

A novel approach for dense reconstruction from a single-view of a repetitive scene structure is presented in this approach [8]. Given an image and its detected repetition regions, we model the shape recovery as the dense pixel correspondences within a single image. The correspondences are represented by an interval map that tells the distance of each pixel to its matched pixels within the single image. In order to obtain dense repetitive structures, A new repetition constraint that penalizes the inconsistency between the repetition intervals of the dynamically corresponding pixel pairs is developed. A graph-cut to balance between the high-level constraint of geometric repetition and the low-level constraints of photometric consistency and spatial smoothness is deployed. The accurate reconstruction of dense 3D repetitive structures through a variety of experiments, which prove the robustness of our approach to outliers such as structure variations, illumination changes, and occlusions is demonstrated.

The existence of repetitive and symmetric structures is a pervasive phenomenon in urban scenes. In typical images, the perspective distorted repetition and symmetry encode the relative 3D geometry between the repeating elements. If the repetition is mostly on a plane, the perspective distortion can be modeled by a planar homography. Detecting such planar repetition and symmetry allows us to recover vanishing points and camera calibrations. While non-planar repeating structures cannot be accurately modeled by one homography, this approach exploits the visual differences between the repeating elements to recover the 3D details. The goal is to obtain the 3D repetition information by modeling it as energy minimization yielding a dense 3D reconstruction of the repetitive structures.

The main contribution of this approach is a novel model to use high-level geometric information, such as repetition and reflective symmetry, in an optimization framework, which allows to enforce geometric consistency between repetitive pixels that are not immediate image neighbors. By enforcing consistency between the 3D reconstruction of the repeating elements, more accurate reconstructions even filling in occluded structures can be achieved. One application of this approach is stereo reconstruction of urban scenes with repetitive 3D structures. Additionally, the extension of the proposed concept to multi-view reconstruction is demonstrated.

I. Consistent Depth Maps Recovery from a Video Sequence

This approach presents a novel method for recovering consistent depth maps from a video sequence [9]. A bundle optimization framework to address the major difficulties in stereo reconstruction is proposed, such as dealing with image noise, occlusions, and outliers. Different from the typical multiview stereo methods, this approach not only imposes the photo-consistency constraint, but also explicitly associates the geometric coherence with multiple frames in a statistical way. It thus can naturally maintain the temporal coherence of the recovered dense depth maps without oversmoothing. To make the inference tractable, we introduce an iterative optimization scheme by first initializing the disparity maps using segmentation prior and then refining the disparities by means of bundle optimization. Instead of defining the visibility parameters, this method implicitly models the reconstruction noise as well as the probabilistic visibility. After bundle optimization, an efficient space-time fusion algorithm to further reduce the reconstruction noise is introduced. The automatic depth recovery is evaluated using a variety of challenging video examples.

Stero reconstruction of dense depth maps from natural video sequences is a fundamentally important and challenging problem in computer vision. The reconstructed depths usually serve as a valuable source of information, and facilitate applications in various fields, including 3D modeling, layer separation, image-based rendering, and video editing. Although the stereo matching problem has been extensively studied during the past decades, automatically computing high-quality dense depths is still difficult on account of the influence of image noise, textureless regions, and occlusions that are inherent in the captured image/video data.

Given an input video sequence taken by a freely moving camera, a novel method to automatically construct a view-dependent depth map for each frame with the following two objectives is proposed. One is to make the corresponding depth values in multiple frames consistent. The other goal is to assign distinctive depth values for pixels that fall in different depth layers. To accomplish these goals, this approach contributes a global optimization scheme, which we call bundle optimization, to resolve most of the aforementioned difficulties in disparity estimation. This framework of allows to produce sharp and temporal consistent object boundaries among different frames.

This method does not explicitly model the binary visibility (or occlusion). Instead, it is encoded naturally in a statistical way with our energy definition. This model also does not distinguish among image noise, occlusions, and estimation outliers, so as to achieve a unified framework for modeling the matching ambiguities. The photo-consistency and geometric coherence constraints associating different views are combined in a global energy minimization framework. They help reliably reduce the influence of image noise and occlusions with the multiframe data, and consequently, make our optimization free from the oversmoothing or blending artifacts.

In order to get an accurate disparity estimate in the textureless region and reduce the problem of false segmentation especially for the fine object structures, approach confine the effect of color segmentation only in the disparity initialization step. Then, the iterative optimization algorithm refines the disparities in a pixelwise manner.

J. Single Image Depth Estimation from Predicted Semantic Labels

In this approach, the problem of estimating the depth of each pixel in a scene from a single monocular image is considered [10]. Unlike traditional approaches, which attempt to map from appearance features to depth directly, this method first performs a semantic segmentation of the scene and use the semantic labels to guide the 3D reconstruction. This approach provides several advantages: By knowing the semantic class of a pixel or region, depth and geometry constraints can be easily enforced (e.g., “sky” is far away and “ground” is horizontal). In addition, depth can be more readily predicted by measuring the difference in appearance with respect to a given semantic class. For example, a tree will have more uniform appearance in the distance than it does close up.

Producing spatially plausible 3D reconstructions of a scene from monocular images annotated with geometric cues (such as horizon, vanishing points, and surface boundaries) is a well understood problem. However, to uniquely determine absolute depths, additional information such as texture, relative depth, and camera parameters (pose and focal length) is needed. Much recent work on automated 3D scene reconstruction has focuses on extracting these geometric cues and additional information from novel images.

In this topic, a different approach that reasons about the semantic content of a scene and uses this information as context for depth reconstruction is proposed. The incorporation of semantic class knowledge allows to do two things: First, the advantage of class-related depth and geometry priors. For example, sky is always at the farthest depth possible; grass and road form supporting ground planes for other objects. Second is by conditioning on semantic class, a better model depth as a

function of local pixel appearance. For example, uniformity of texture may be a good indicator for the depth of a tree, but not useful when estimating the depth of a building.

K. Depth estimation from a video sequence with moving and deformable objects

In this approach, an algorithm for depth estimation from a monocular video sequence containing moving and deformable objects is presented [11]. The method is based on a coded aperture system (i.e., a conventional camera with a mask placed on the main lens) and it takes a coded video as input to provide a sequence of dense depth maps as output. To deal with nonrigid deformations, the work builds on the state-of-the art single-image depth estimation algorithm. Since single image depth estimation is very ill-posed, the reconstruction task as a regularized algorithm based on nonlocal means filtering applied to both the spatial and temporal domain is casted. The assumption is that regions with similar texture in the same frame and in neighbouring frames are likely to belong to the same surface. Moreover, it shows how to increase the computational efficiency of the method. The proposed algorithm has been successfully tested on challenging real scenarios.

This work provides the following three main contributions:

- It presents, to the best of our knowledge, the first single-frame video depth estimation algorithm, capable of handling moving and deformable objects;
- It introduces a novel spatial and temporal depth smoothness constraint, based on nonlocal-means (NLM) filtering: Pixels whose intensities match within a certain spatial and temporal range are likely to share similar depths;
- The proposed algorithm is robust and accurate on real videos.

CONCLUSION AND FUTURE WORK

So far, from the number of approaches studied we can conclude that the automatic techniques are applicable to estimate depths for videos. Methods are also applicable in cases where other methods fail, such as those based on motion parallax and structure from motion, and works even for single images and dynamics scenes, and our single-image algorithm quantitatively outperforms existing methods. Using 3D reconstruction techniques, we can generate stereoscopic videos for 3D viewing from conventional 2D videos. A survey on Depth map extraction is also suitable as a good starting point for converting legacy 2D feature films into 3D.

REFERENCES

[1] G. Zhang, J. Jia, W. Hua, and H. Bao, "*Robust bilayer segmentation and motion/depth estimation with a handheld camera*," IEEE TPAMI, vol. 33, no. 3, pp. 603–617, 2011.
[2] Y. Horry, K. Anjyo, and K. Arai, "*Tour into the picture: Using a spidery mesh interface to make animation from a single image*," SIGGRAPH, 1997.
[3] B. Oh, M. Chen, J. Dorsey, and F. Durand, "*Image-based modeling and photo editing*," SIGGRAPH, 2001.
[4] D. Hoiem, A. Efros, and M. Hebert, "*Automatic photo popup*," in ACM SIGGRAPH, 2005.
[5] E. Delage, H. Lee, and A. Ng, "*A dynamic Bayesian network model for autonomous 3D reconstruction from a single indoor image*," in CVPR, 2006.
[6] A. Saxena, S. Chung, and A. Ng, "*Learning depth from single monocular images*," NIPS, 2005.
[7] A. Saxena, M. Sun, and A. Ng, "*Make3D: Learning 3D scene structure from a single still image*," IEEE TPAMI, vol. 31, no. 5, pp. 824–840, 2009.
[8] C.Wu, J.-M. Frahm, and M. Pollefeys, "*Repetition-based dense single-view reconstruction*," CVPR, 2011.
[9] G. Zhang, J. Jia, T.-T. Wong, and H. Bao, "*Consistent depth maps recovery from a video sequence*," IEEE TPAMI, vol. 31, pp. 974–988, 2009.
[10] B. Liu, S. Gould, and D. Koller, "*Single image depth estimation from predicted semantic labels*," in CVPR, 2010.
[11] Martinello, Manuel; Favaro, Paolo, "*Depth estimation from a video sequence with moving and deformable objects*," Image Processing (IPR 2012), IET Conference on , vol., no., pp.1,6, 3-4 July 2012.

