ISSN: 2321-9653

INTERNATIONAL JOURNAL FOR RESEARCH IN APPLIED SCIENCE AND ENGINEERING TECHNOLOGY (IJRASET)

Recent Trends in 2D to 3D Conversion: A Survey

Jean Maria Dominic

Computer Science Department, Viswajyothi College of Engineering and Technology

MG University, Kerala

Abstract— The concept of stereoscopy has existed for a long time. But the breakthrough from conventional 2D broadcasting to real-time 3D broadcasting is still pending. However, in recent years, there has been rapid progress in the field's image capture, coding and display which brings the realm of 3D closer to reality than ever before. The survey investigates the existing 2D to 3D conversion algorithms developed in the past years by various computer vision research communities across the world. Each algorithm has its own strengths and weaknesses. Most conversion algorithms make use of certain depth cues to generate depth maps. Among 2D-to-3D image conversion methods, those involving human operators have been most successful but also time-consuming and costly. Fully-automatic methods typically make strong assumptions about the 3D scene. Although such methods may work well in some cases, in general it is very difficult to construct a deterministic scene model that covers all possible background and foreground combinations. In practice, such methods have not achieved the same level of quality as the semi-automatic methods.

Keywords— 3D images, stereoscopic images, image conversion, depth estimation, depth maps.

I. INTRODUCTION

Stereoscopy, also called stereoscopic or 3D imaging is a technique for creating or enhancing the illusion of depth in an image by means of stereopsis for binocular vision. Most stereoscopic methods present two set images separately to the left and right eye of the viewer. These two-dimensional images are then combined in the brain to give the perception of 3D depth. This technique is distinguished from 3D displays that display an image in three full dimensions, allowing the observer to increase information about the 3-dimensional objects being displayed by head and eye movements.

Three-dimensional television (3D-TV) is nowadays often seen as the next major milestone in the ultimate visual experience of media. Although the concept of stereoscopy has existed for a long time, the breakthrough from conventional 2D broadcasting to real-time 3D broadcasting is still pending. However, in recent years, there has been rapid progress in the fields image capture, coding and display, which brings the realm of 3D closer to reality than ever before.

The world of 3D incorporates the third dimension of depth, which can be perceived by the human vision in the form of binocular disparity. Human eyes are located at slightly different positions, and these perceive different views of the real world. The brain is then able to reconstruct the depth information from these different views. A 3D display takes advantage of this phenomenon, creating two slightly different images of every scene and then presenting them to the individual eyes. With an appropriate disparity and calibration of parameters, a correct 3D perception can be realized.

An important step in any 3D system is the 3D content generation. Several special cameras have been designed to generate 3D model directly. For example, a stereoscopic dual-camera makes use of a co-planar configuration of two separate, monoscopic cameras, each capturing one eye's view, and depth information is computed using binocular disparity. A depth-range camera is another example. It is a conventional video camera enhanced with an add-on laser element, which captures a normal two-dimensional RGB image and a corresponding depth map. A depth map is a 2D function that gives the depth of an object point as a function of the image

ISSN: 2321-9653

coordinates. Usually, it is represented as a grey level image with the intensity of each pixel registering its depth. The laser element emits a light wall towards the real world scene, which hits the objects in the scene and reflected back. This is subsequently registered and used for the construction of a depth map.

II. State of the Art

Two approaches to 2D to 3D conversion can be loosely defined: quality semiautomatic conversion for cinema and high quality 3DTV, and low-quality automatic conversion for cheap 3DTV, VOD and similar applications. In semiautomatic conversion a skilled operator assigns depth to various parts of an image or video. Based on this sparse depth assignment, a computer algorithm estimates dense depth over the entire image or video sequence. In the case of automatic methods, no operator intervention is needed and a computer algorithm automatically estimates the depth for a single image or video. Automatic methods estimates shape from shading, structure from motion or depth from defocus. Electronics manufacturers use stronger assumptions to develop real-time 2D-to-3D converters. Such methods may work well in specific scenarios. But generally it is very difficult to construct heuristic assumptions that cover all possible background and foreground combinations.

In order to reduce operator involvement in the semiautomatic conversion process and therefore, lower the cost while speeding up the conversion, research effort has recently focused on the most labour-intensive steps of the manual involvement, namely spatial depth assignment. Guttmann et al. [4] have proposed a dense depth recovery via diffusion from sparse depth assigned by the operator. The focus of the method proposed by Agnot et al. [7] is the application of cross-bilateral filtering to an initial depth map. The authors propose to use a library of initial depth maps from which an operator can choose one that best corresponds to the image being converted. They also suggest estimation of the initial depth map based on image blur but show only one very simple example; this initialization is unlikely to work well in more complex cases. Phan et al. [12] propose a simplified and more efficient version of the Guttmann et al. [4] method using scale-space random walks that they solve with the help of graph cuts. Liao et al.[9] further simplify operator involvement by first computing optical flow, then applying structure-from-motion estimation and finally extracting moving object boundaries. The role of an operator is to correct errors in the automatically computed depth of moving objects and assign depth in undefined areas.

The problem of depth estimation from a single 2D image, which is the main step in 2D-to-3D conversion, can be formulated in various ways, for example as a shape-from shading problem [16]. However, this problem is severely under-constrained; quality depth estimates can be found only for special cases. Other methods, often called multi-view stereo, attempt to recover depth by estimating scene geometry from multiple images not taken simultaneously. For example, a moving camera permits structure-from-motion estimation [18] while a fixed camera with varying focal length permits depth from-defocus estimation [19]. Both are examples of the use of multiple images of the same scene captured at different times or under different exposure conditions. Although such methods are similar in spirit to the methods proposed here, the main difference is that while these methods use images known to depict the same scene as the query image, the proposed method uses all images available in a large repository and automatically select suitable ones for depth recovery.

Recently, machine-learning-inspired techniques employing image parsing have been used to estimate the depth map of a single monocular image [5], [3]. Such methods have the potential to automatically generate depth maps, but currently work only on few types of images using carefully-selected training data. The data-driven approaches to 2D-to-3D conversion is inspired by the recent trend to use large image databases for various computer vision tasks, such as object recognition [14] and image saliency detection [15].

A detailed study of few 2D to 3D conversion methods are given below:

A. A Semi-Automatic 2D to 3D Image Conversion Using Scale-Space Random Walks And A Graph Cuts Based Depth Prior

Here, a semi-automated method is used for converting conventional 2D images into stereoscopic 3D. User-defined strokes corresponding to a rough estimate of the depth values

ISSN: 2321-9653

in the scene are defined for the image of interest. With these, system determines the depth values for the rest of the image, producing a depth map that can be used to create stereoscopic 3D image pairs. The work is based on a similar scheme, using the Random Walks segmentation paradigm. However, the related work is quite complex, with many processing steps required to produce the final stereoscopic image pair. Combined with its evident shortcomings, but noting the merits, a system employing Random Walks is proposed, while incorporating information from the popular Graph Cuts segmentation paradigm. Thus, a final cohesive depth map is produced, combining the merits of both. The results show that the project produces good quality stereoscopic image pairs, while using a much more simplified method in comparison to the related work.

Fig. 1 The Cabot Tower example along with associated labelling and depth maps (a) Labelled Image (b) Graph Cuts (c) Random Walks.

Generating depth maps in a segmentation-based framework is an intuitive process. Rather than just considering each label as a separate object, here each label is considered as a separate depth, and can ultimately be seen as a case of multi-label segmentation. The user merely has to mark each object and specify their relative depths. This is sufficient, as noted, the exact depth values do not have to be known. This is a two stage process using the smoothing properties of Random Walks and the hard segmentation returned by Graph Cuts. Random Walks is the solution to a linear system and has problems preserving strong edges, but Graph Cuts does this quite well. However, the hard segmentation with Graph Cuts does not respect smooth gradients or fine detail. By combining the two, we can retain strong object boundaries while also allowing for smooth gradients. There has already been work that has merged the merits of the two in a unified segmentation framework. An initial depth map using Graph Cuts is generated first with user-defined depth strokes, in order to generate a depth prior. The depth prior and the same depth strokes are integrated into Random Walks as an additional feature when determining the edge weights. The merits of Random Walks are combined with Graph Cuts, in order to produce an augmented, good quality depth map.

1) Advantages: The results show that this method produces good quality stereoscopic image pairs. A much more simplified method is used in comparison to the related work.

2) Disadvantages: Human interactions are required. Also 2D to 3D video conversion is not possible.

B. Auto-Directed Video Stabilization with Robust L1 Optimal Camera Paths

A novel algorithm for automatically applying constrainable, an L1-optimal camera path to generate stabilized videos by removing undesired motions is presented here. The goal is to compute camera paths that are composed of constant, linear and parabolic segments mimicking the camera motions employed by professional cinematographers. To this end, the algorithm is based on a linear programming framework to minimize the first, second, and third derivatives of the resulting camera path. This method allows for video stabilization beyond the conventional filtering of camera paths that only suppresses high frequency jitter. Additional constraints are incorporated on the path of the camera directly in the algorithm, allowing for stabilized and retargeted videos. The approach presented here accomplishes this without the need of user interaction or costly 3D reconstruction of the scene, and works as a post-process for videos from any camera or from an online source. This technique may not be able to stabilize all videos. Example, Low feature count, excessive blur during extremely fast motions Lack of rigid objects in the scene might make camera path estimation unreliable. The use of cropping discards information, something a viewer might dislike.

ISSN: 2321-9653

This algorithm works as a post process and can be applied to videos from any camera or from an online source without any knowledge of the capturing device or the scene. A post-process video stabilization consists of the following three main steps: (1) Estimating the original camera path, (2) Estimating a new smooth camera path, and (3) Synthesizing the stabilized video using the estimated smooth camera path. The key contribution of a method is a novel algorithm to compute the optimal steady camera path. A crop window is moved of fixed aspect ratio along this path; a path optimized to include salient points and regions, while minimizing an L1-smoothness constraint based on cinematography principles. This technique finds optimal partitions of smooth paths by breaking the path into segments of constant, linear, or parabolic motion. It avoids the superposition of these three types, resulting in, for instance, a path that is truly static within a constant segment instead of having small residual motions. Furthermore, it removes low-frequency bounces, e.g. those originating from a person walking with a camera. The optimization is poses as a Linear Program (LP) subject to various constraints, such as inclusion of the crop window within the frame rectangle at all times. Any additional motion inpainting is not performed.

Fig. 2 Example from YouTube "Fan-Cam" video, Top row: Stabilized rebottom row: Original with optimal crop window.

1) Advantages: The video stabilization beyond the conventional filtering of camera paths that only suppresses high frequency jitter is used. User interaction or costly 3D reconstruction of the scene is not required. It works as a postprocess for videos from any camera or from an online source.

2) Disadvantages: This technique may not be able to stabilize all videos. Lack of rigid objects in the scene might make camera path estimation unreliable.The use of cropping discards information, something a viewer might dislike.

C. Depth Extraction from Video Using Non-parametric Sampling

This method describes a technique that automatically generates plausible depth maps from videos using non-parametric depth sampling. This technique is demonstrated in cases where past methods fail (nontranslating cameras and dynamic scenes). This technique is applicable to single images as well as videos. For videos, local motion cues are used to improve the inferred depth maps, while optical flow is used to ensure temporal depth consistency. For training and evaluation, a Kinect-based system is used to collect a large dataset containing stereoscopic videos with known depths.

From a given input image, find matching candidates in the database, and warp the candidates to match the structure of the input image. Then use a global optimization procedure to interpolate the warped candidates producing per-pixel depth estimates for the input image. With temporal information (e.g. extracted from a video), this algorithm can achieve more accurate, temporally coherent depth. The depth estimation technique used in this method outperforms the state-of-the-art on benchmark databases. This technique can be used to automatically convert a monoscopic video into stereo for 3D visualization, and this is demonstrate through a variety of visually pleasing results for indoor and outdoor scenes, including results from the feature film Charade.

1) Advantages: Automatically convert a monoscopic video into stereo for 3D visualization. Nontranslating cameras and dynamic scenes cases can be demonstrated using this. Also, this method is applicable to single images as well as videos.

2) Disadvantages: In some cases, motion segmentation misses or falsely identifies moving pixels. This can result in inaccurate depth and 3D estimation. The algorithm also assumes that moving objects contact the ground, and thus may

ISSN: 2321-9653

fail for airborne objects. Due to the serial nature of this method is prone to propagating errors through the stages. For example, if an error is made during depth estimation, the result may be visually implausible.

D. Video Stereolization: Combining Motion Analysis with User Interaction

This method present a semi-automatic system that converts conventional videos into stereoscopic videos by combining motion analysis with user interaction, aiming to transfer as much as possible labeling work from the user to the computer. In addition to the widely-used structure from motion (SFM) techniques, two new methods are developed that analyze the optical flow to provide additional qualitative depth constraints. They remove the camera movement restriction imposed by SFM so that general motions can be used in scene depth estimation – the central problem in mono-to-stereo conversion. With these algorithms, the user's labelling task is significantly simplified. A quadratic programming approach is also developed to incorporate both quantitative depth and qualitative depth (such as these from user scribbling) to recover dense depth maps for all frames, from which stereoscopic view can be synthesized. The user study results show that this approach is more intuitive and less labour intensive, while producing 3D effect comparable to that from current interactive algorithms.

In the pre-processing step, the input image sequence is first passed through three individual automatic modules: structure-from-motion (SFM) moving object segmentation (MOS), and perspective depth correction (PDC). The SFM algorithm is applied to the input image sequence with dominant rigidly moving objects to recover a sparse set of 3D points. The MOS module is used to automatically segment the foreground, it is particularly effective in a follow shot in which the foreground is relatively static and the background is rapidly changing. Finally, the PDC module inspects the size change of an object's image to estimate relative depth changes between frames. After automatic processing, the users are presented with images showing area with known depth (from SFM and MOS). If there are still undefined regions, the users need to label them in some key frames by simple scribbling. The user's input as well as all the automatically calculated depth cues will be integrated in a quadratic programming framework to generate dense depth maps for all frames. Finally the novel view is generated via shifting every pixel horizontally by a certain amount base on the depth maps, simulating the perspective from the other eye. Since in most of the cases, the baseline between the synthesized view and the input view is small, a simple technique is used to deal with the gaps in the synthesized view due to disocclusion. Fill the uncoloured region with neighbouring pixels of larger depth values.

1) Advantages: The major novelty of the framework lies in the utilization of motion prior analysis.User interface requires users to specify relative depth orders with the help of pre-computed 3D visual cues, instead of labelling the depth value directly.

2) Disadvantages: The success of the automatic processing modules depends on the camera/object motion. If there is no feature to track, the resulting depth map will not be accurate.

E. 2D-to-3D Image Conversion by Learning Depth from Examples

Among 2D-to-3D image conversion methods, those involving human operators have been most successful but also time-consuming and costly. Automatic methods, that typically make use of a deterministic 3D scene model, have not yet achieved the same level of quality as they often rely on assumptions that are easily violated in practice. Here, radically different approach of "learning" the 3D scene structure is adopted. A simplified and computationally-efficient version of 2D-to-3D image conversion algorithm is developed. From a given repository of 3D images, either as stereopairs or image+depth pairs, find k pairs whose photometric content most closely matches that of a 2D query to be converted. Then, fuse the k corresponding depth fields and align the fused depth with the 2D query. The simplified algorithm validated quantitatively on a Kinect-captured image+depth dataset against the Make3D algorithm. While far from perfect, the presented results demonstrate that online repositories of 3D content can be used for effective 2Dto-3D image conversion.

This approach is built upon a key observation and an assumption. The key observation is that among millions of image+depth pairs available on-line, there likely exist many

pairs whose 3D content matches that of a 2D input query. The assumption is that two images that are photometrically similar are likely to have similar 3D structure depth. This is not unreasonable since photometric properties are often correlated with 3D content. For example, edges in a depth map almost always coincide with photometric edges.

The proposed algorithm compares favorably in terms of both estimated depth quality and computational complexity. Admittedly, the validation was limited to a database of indoor scenes on which Make3D was not trained. The generated anaglyph images produce a comfortable 3D perception but are not completely void of distortions.

1) Advantages: While far from perfect, the presented results demonstrate that online repositories of 3D content can be used for effective 2dto-3d image conversion. This method is favourable in terms of both estimated depth quality and computational complexity.

2) Disadvantages: Uses SIFT which bring additional computation complexity. The validation was limited to a database of indoor a scene on which make3d was not trained. The generated anaglyph images produce a comfortable 3d perception but are not completely void of distortions.

F. Learning-Based, Automatic 2D-to-3D Image And Video Conversion

Among 2D-to-3D image conversion methods, those involving human operators have been most successful but also time-consuming and costly. Fully-automatic methods typically make strong assumptions about the 3D scene. Although such methods may work well in some cases, in general it is very difficult to construct a deterministic scene model that covers all possible background and foreground combinations. In practice, such methods have not achieved the same level of quality as the semi-automatic methods. Two types of methods are used in the project. The first one is based on learning a point mapping from local image/video attributes, such as color, spatial position, and motion at each pixel, to scene-depth at that pixel using a regression type idea.

The second one is based on globally estimating the entire depth map of a query image directly from a repository of 3D images which is a set of image + depth pairs using a nearest-neighbor regression type idea.This approach is built upon a key observation and an assumption. The key observation is that among millions of 3D images available on-line, there likely exist many whose 3D content matches that of the 2D input query. The key assumption is that two 3D images whose left images are photometrically similar are likely to have similar depth fields.

Fig. 3 Anaglyph images generated using the ground-truth depth (a) and depths estimated by the proposed global method (b).

1) Advantages: The 2d-to-3d conversion based on learning a local point transformation has the undisputed advantage of computational efficiency. The point transformation can be learned off-line and applied basically in real time. The same transformation can be applied to images with potentially different global 3d scene structure. For global method millions of 3D images are available on-line. It requires less computation time only. Also, it has reduced complexity compared to the previous methods.

2) Disadvantage: Low resolution images give better output.

III. CONCLUSIONS

The survey investigates the existing 2D to 3D conversion algorithms developed in the past years by various computer vision research communities across the world. The results of some 2D to 3D conversion algorithms are 3D coordinates of a small set of points in the images. This group of algorithms is less suitable for the 3D television application. The depth cues based on multiple images yield in general more accurate

results, while the depth cues based on single still image are more versatile. A single solution to convert the entire class of 2D images to 3D models does not exist. Combing depth cues enhances the accuracy of the results. It has been observed that machine learning is a new and promising research direction in 2D to 3D conversion. It is also helpful to explore the alternatives than to confine ourselves only in the conventional methods based on depth maps.

ACKNOWLEDGMENT

The author would like to express her sincere thanks to HOD, group tutor and staff in Computer Science department, Viswa Jyothi College of Engineering and Technology for many fruitful discussions and constructive suggestions during the preparation of this manuscript.

REFERENCES

[1] Janusz Konrad, Fellow, Meng Wang, Prakash Ishwar,Chen Wu, and Debargha Mukherjee "Learning-Based, Automatic 2D-to-3D Image and Video Conversion",IEEE Trans. Image Process., vol. 22, no. 9, pp. 3485_3496, Sep. 2013.

[2] A. Saxena, S. H. Chung, and A. Y. Ng, "Learning depth from single monocular images," in Advances in Neural Information Processing Systems. Cambridge, MA,USA: MIT Press, 2005.

[3] A. Saxena, M. Sun, and A. Ng, _Make3D: Learning 3D scene structure from a single still image,_ IEEE Trans. Pattern Anal. Mach. Intell.,vol. 31, no. 5, pp. 824_840, May 2009.

[4] M. Guttmann, L. Wolf, and D. Cohen-Or, "Semi-automatic stereo extraction from video footage", in Proc. IEEE Int. Conf. Comput. Vis., Oct. 2009, pp. 136_142.

[5] B. Liu, S. Gould, and D. Koller, "Single image depth estimation from predicted semantic labels," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2010, pp. 1253_1260.

[6] R. Phan, R. Rzeszutek, and D. Androutsos, "Semi-automatic 2D to 3D image conversion using scale-space random walks and a graph cuts based depth prior,"in Proc. 18th IEEE Int. Conf. Image Process., Sep. 2011, pp. 865_868.

[7] M. Grundmann, V. Kwatra, and I. Essa, "Auto-directed video stabilization with robust L1 optimal camera paths," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.,Jun. 2011, pp. 225_232.

[8] K. Karsch, C. Liu, and S. B. Kang, "Depth extraction from video using non- parametric sampling," in Proc. Eur. Conf. Comput. Vis., 2012, pp. 775_788.

[9] M. Liao, J. Gao, R. Yang, and M. Gong, "Video stereolization: Combining motion analysis with user interaction," IEEE Trans. Visualizat. Comput. Graph., vol. 18, no. 7, pp. 1079_1088, Jul. 2012.

[10] J. Konrad, M. Wang, and P. Ishwar, "2D-to-3D image conversion by learning depth from examples," in Proc. IEEE Comput. Soc. CVPRW, Jun. 2012, pp. 16_22.

[11] Ming-Fu Hung, Shaou-Gang Miaou, and Chih-Yuan Chiang "Dual Edge-Con_ned Inpainting of 3D Depth Map Using Color Image's Edges and Depth Image's Edges" Signal and Information Processing Association Annual Summit and Conference (AP-SIPA), 2013 Asia-Pacific Oct. 29 2013-Nov. 1 2013.

[12] J. Konrad, G. Brown, M. Wang, P. Ishwar, C. Wu, and D. Mukherjee, "Automatic 2D-to-3D image conversion using 3D examples from the Internet,"Proc. SPIE, vol. 8288, p. 82880F, Jan. 2012.

[13] L. Angot, W.-J. Huang, and K.-C. Liu, "A 2D to 3D video and image conversion technique based on a bilateral" lter,_ Proc. SPIE, vol. 7526, p. 75260D, Feb. 2010.

[14] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, no. 11, pp. 1958_1970, Nov. 2008.

[15] M. Wang, J. Konrad, P. Ishwar, K. Jing, and H. Rowley, "Image saliency: From intrinsic to extrinsic context," in

Proc. IEEE Conf. Comput. Vis. Pattern Recognit.,Jun. 2011, pp. 417_424.

[16] R. Zhang, P. S. Tsai, J. Cryer, and M. Shah, "Shape-from-shading: A survey," IEEE Trans. Pattern Anal. Mach. Intell., vol. 21, no. 8, pp. 690_706, Aug. 1999.

[17] L. Angot, W.-J. Huang, and K.-C. Liu, "A 2D to 3D video and image conversion technique based on a bilateral" lter,_ Proc. SPIE, vol. 7526, p. 75260D, Feb. 2010.

[18] R. Szeliski and P. H. S. Torr, "Geometrically constrained structure from motion: Points on planes," in Proc. Eur. Workshop 3D Struct. Multiple Images Large-Scale Environ., 1998, pp. 171_186.

[19] M. Subbarao and G. Surya, "Depth from defocus: A spatial domain approach," Int.J. Comput. Vis., vol. 13, no. 3, pp. 271_294, 1994.