Automatic 2D to 3D Video Conversion based on Global Nearest Neighbor Depth Learning

Ms. Baby Shamna. P
Electronics and Communication
KMCT College of Engineering
Kallanthode, Calicut, India

Mr. Aji George
Electronics and Communication
KMCT College of Engineering
Kallanthode, Calicut, India

***Abstract*- The conversion of existing 2D videos to 3D is proving commercially viable and fulfills the growing need for high quality stereoscopic videos. Most successful 2D-to-3D image and video conversion methods involve human operators, which is time-consuming and costly. Automatic methods, which typically make use of a deterministic 3D scene model, have not yet achieved the same level of quality for they rely on assumptions that are often violated in practice. In this paper, there is a new method that is based on the radically different approach of learning the 2D-to-3D video conversion from examples. The method is based on globally estimating the entire depth map of each frame that we are selected for the conversion. The approach proposes here is built upon a key observation and an assumption. The key observation is that among millions of 3D frames (images) available on-line, there likely exist many whose 3D content matches that of a 2D input (query) we wish to convert to 3D. We are also making an assumption that two frames that are photo metrically similar also have similar 3D structure (depth).**

Keywords*— 3D frames, stereoscopic video, nearest neighbor classification, cross-bilateral filtering.

I. INTRODUCTION

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 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 or video 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 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.

The last few years have seen a dramatic increase in the demand for stereo content. This has largely been driven by the commercial availability of multi viewer auto stereoscopic displays. Stereoscopy also called stereoscopic or is a technique for creating or enhancing the illusion of depth in an image or video by means of stereopsis for binocular vision.

There are two basic approaches to 2D-to-3D conversion: one That requires a human operator's intervention and one that does not. In the former case, the so-called semi-automatic methods have been proposed where 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. The involvement of a human operator may vary from just a few scribbles to assign depth to various locations in an image or frames to a precise delineation of objects and subsequent depth assignment to the delineated regions. 3D capable hardware like 3D TVs, Blu-Ray players, handheld gaming consoles, cell phones, still and video cameras are widely available in the market. But this hardware availability is not yet matched 3D content production. Current available 2D to 3D conversion methods has not achieved a high quality level. The most successful approaches are interactive. That means it involve human operators.

The method proposes in this paper; carry the "big data" philosophy of machine learning. In consequence, they apply to arbitrary scenes and require no manual annotation. The data driven approach to 2D-to-3D conversion has been inspired by the recent trend to use large image databases for various computer vision tasks, such as object recognition [8] and image saliency detection. In particular, this proposes a new method that is based on the radically different approach of learning the 2D-to-3D conversion from examples. The method is based on globally estimating the entire depth map of a query frame directly from a repository of 3D frames or images (frame+ depth pairs or stereo pairs) using a nearest-neighbor regression type idea. Early versions of learning-based approach to 2D-to-3D image or video conversion, suffered from high computational complexity .The method demonstrate the improved quality of the depth maps produced by our global method relative to state of- the-art methods together with up to 4 orders of magnitude reduction in computational effort.

While 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 is applied to images or frames with potentially different global 3D scene structure. This is because this type of conversion, although learning-based, is based on purely local image/video attributes, such as color, spatial position, and motion at each pixel. To address this limitation, in this section which develop a method that estimates the global depth map of a query image or video frame directly from a repository of 3D images or frames (image(frames)+depth pairs or stereo pairs) using a nearest-neighbor regression type idea.

II. STATE OF THE ART

There are two types of 2D-to-3D video conversion methods: Semi-automatic methods, that require human operator intervention, and automatic methods, that require no such help.

A. Semi-Automatic Methods

To date, this has been the more successful approach to 2D to -3D conversion. In fact, methods that require a significant operator intervention in the conversion process, such as delineating objects in individual frames, placing them at suitable depths, and correcting errors after final rendering, have been successfully used commercially by such companies as IMAX Corp., Digital Domain Productions Inc. (formerly In-Three Inc.), etc. Many films have been converted to 3D using this approach. In order to reduce operator involvement in the process and, therefore, lower the cost while speeding up the conversion, research effort has recently focused on the most labor-intensive steps of the manual involvement, namely spatial depth assignment. Guttmann *et al.* [6] have proposed a dense depth recovery *via* diffusion from sparse depth assigned by the operator. In the first step, the operator assigns relative depth to image patches in some frames by scribbling. In the second step, a combination of depth diffusion, which accounts for local image saliency and local motion, and depth classification, is applied. In the final step, disparity is computed from the depth field and two novel views are generated by applying half of the disparity amplitude. The focus of the method proposed by Agnot *et al.* [2] is the application of cross-bilateral filtering to an initial depth map. The authors propose to use a library of initial depth maps (smooth maps consistent with the 3D perspective of outdoor scenes or rooms) from which an operator can choose one that best corresponds to the frame being converted.

They also suggest estimation of the initial depth map based on image or frame blur but show only one very simple example; this initialization is unlikely to work well in more complex cases. Phan *et al.* [7] propose a simplified and more efficient version of the Guttmann *et al.* [6] 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.

B. Automatic Methods

The problem of depth estimation from a single 2D image or video, which is the main step in 2D-to-3D conversion, can be Formulated in various ways, for example as a shape-from shading problem. 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 frames not taken simultaneously. For example, a moving camera permits structure-from-motion estimation [10] while a fixed camera with varying focal length permits depth from- defocus estimation [11]. Both are examples of the use of multiple frames of the same video 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 frames known to depict the same scene as the query image, this method use all frames available in a large repository and automatically select suitable ones for depth recovery.

Several electronics manufacturers have developed real-time 2D-to-3D converters that rely on stronger assumptions and simpler processing than the methods discussed above, e.g., faster-moving or larger objects are assumed to be closer to the Viewer, higher frequency of texture is assumed to belong to objects located further away, etc. Although such methods may work well in specific scenarios, in general it is very difficult, if not impossible, to construct heuristic assumptions that cover all possible background and foreground combinations. Such real-time methods have been implemented in Blu-Ray 3D players by LG, Samsung, Sony and others. DDD offers its TriDef 3D software for PCs, TVs and mobile devices. However, these are proprietary systems and no information is available about the assumptions used.

In the very first attempt, developed a method that fuses SIFT-aligned depth maps selected from a large 3D database; however this approach proved to be computationally demanding [3]. Subsequently, we skipped the costly SIFT based depth alignment and used a different metric (based on histogram of gradients) for selecting most similar depth fields from a database. We observed no significant quality degradation but a significant reduction of the computational Complexity .Very recently, Karsch *et al.* [12] have proposed a depth extraction method based on SIFT warping that essentially follows our initial, unnecessarily complex, Very recently, Karsch *et al.* [12] have proposed a depth extraction method based on SIFT warping that essentially follows our initial, unnecessarily complex, approach to depth extraction [3].

III.PROPOSED METHOD

In this section there is a method that estimates the global depth map of a query video frame directly from a repository of 3D frames or images (frame +depth pairs or stereo pairs) using a nearest-neighbor regression type idea. The approach we propose here 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 a 2D input (query) we wish to convert to 3D. Figure.1 represents the block diagram for the proposed method .We are also making an assumption that two frames that are photo metrically similar also have similar 3D structure (depth). This is not unreasonable since photometric properties are often correlated with 3D content (depth, disparity). For example, edges in a depth map almost always coincide with photometric edges. Given a monocular query frame Q,

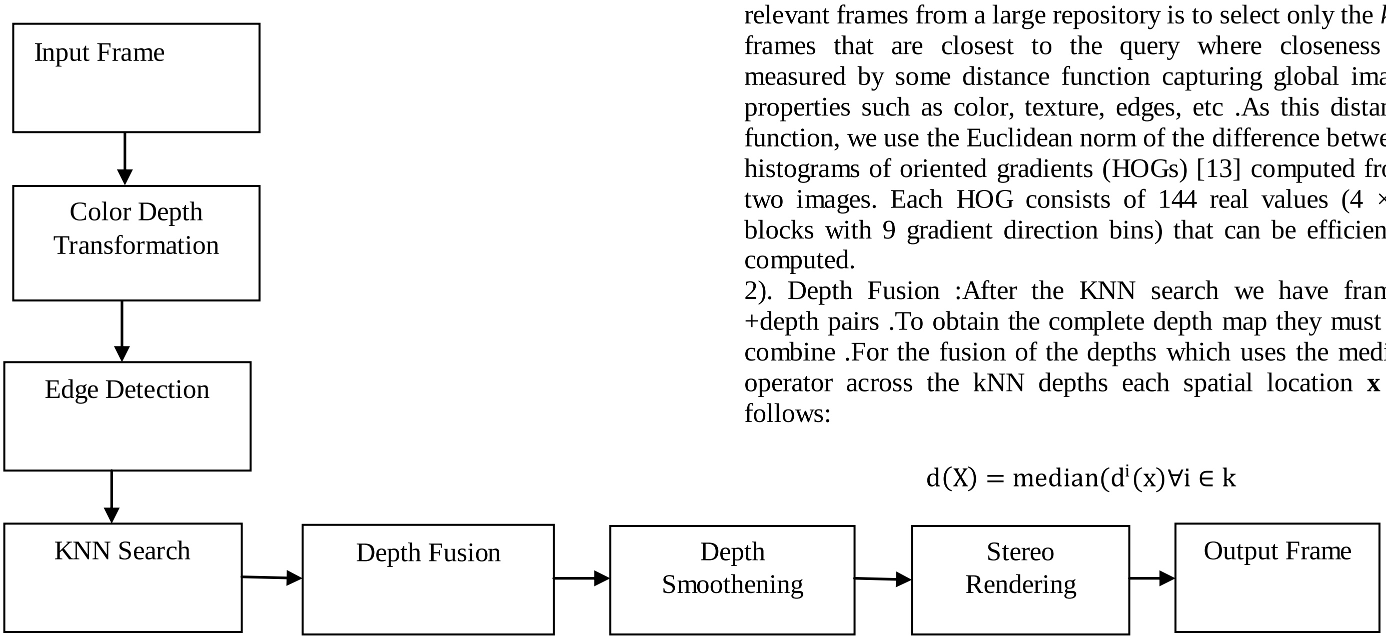

Fig.1.Block Diagram for proposed method

to "learn" the entire depth field from a repository of 3D images or frames and render a stereo pair in the following steps:

LOCAL POINT TRANSFORMATION

1).Color Depth Transformation: In order to obtain a color depth transformation *fc*, we first transform the *YUV* space, commonly used in compressed images and videos, to the *HSV* Color space . We found out that the saturation component (*S*) Provides little depth discrimination capacity and therefore we limit the transformation attributes to hue (*H*) and value (*V*). The depth mapping *fc*[*h*, *v*], *h*, *v* = 1, ..., *L* is computed as the average of depths at all pixels in *I* with hue *h* and value *v*:

$$fc[h,v] = \frac{\sum_{k=1}^{K}\sum_{x} 1(H^k[X] = h, V^k[X] = v)d^k[X]}{\sum_{k=1}^{K}\sum_{x} 1\,(H^k[X] = h, V^k[X] = v)}$$

2).Edge Detection :find out the edges of the each frames .The method utilizes edge information from both depth frame and input frame to find unmatched edge locations .

GLOBAL NEAREST NIGHBOR DEPTH LEARNING

1).KNN Search: There exist two types of images or frames in a large 3D image repository: those that are relevant for determining depth in a 2D query, and those that are irrelevant. Frames that are not photo metrically similar to the 2D query need to be rejected because they are not useful for estimating depth (as per our assumption). Note that although we might miss some depth-relevant frames, we are effectively limiting the number of irrelevant frames that could potentially be more harmful to the 2D-to-3D conversion process. The selection of a smaller subset of frames provides the added practical benefit of computational tractability when the size of the repository is very large. One method for selecting a useful subset of depth-relevant frames from a large repository is to select only the *k* frames that are closest to the query where closeness is measured by some distance function capturing global image properties such as color, texture, edges, etc .As this distance function, we use the Euclidean norm of the difference between histograms of oriented gradients (HOGs) [13] computed from two images. Each HOG consists of 144 real values (4 × 4 blocks with 9 gradient direction bins) that can be efficiently computed.

2). Depth Fusion :After the KNN search we have frames +depth pairs .To obtain the complete depth map they must be combine .For the fusion of the depths which uses the median operator across the kNN depths each spatial location **x** as follows:

$$\mathrm{d(X) = median(d^i(x)\forall i \in k}$$

3). Cross-Bilateral Filtering (CBF) of Depth: While the median-based fusion helps make depth more consistent globally, the fused depth is overly smooth and local inconsistent with the query frame due to edge misalignment between the depth fields of the kNNs and the query frame. This, in turn, often results in the lack of edges in the fused depth where sharp object boundaries should occur and/or the lack of fused-depth smoothness where smooth depth is expected. In order to correct this, similarly to Agnot et al. [1], we apply cross-bilateral filtering (CBF). CBF is a variant of bilateral filtering, an edge-preserving image smoothing method that applies anisotropic diffusion controlled by the local content of the image itself .In CBF, however, the diffusion is not controlled by the local content of the image under smoothing but by an external input. We apply CBF to the fused depth $\hat{d}$ using the query frame Q to control diffusion. This allows us to achieve two goals simultaneously: alignment of the depth edges with those of the luminance Y in the query Frame Q and local noise/granularity suppression in the fused depth $\hat{d}$. This is implemented as follows:

$$\hat{d}(X) = \frac{1}{\gamma(X)}\sum_{y} d[y]h_{\sigma_s}(X - Y)h_{\sigma_e}\,(y(X) - y(Y))$$

$$\gamma(X) = \sum_y \sum_y h_{\sigma_S}(X - Y)h_{\sigma_e}\,(y(X) - y(Y))$$

Where, $\hat{d}$ is the filtered depth field and $h_\sigma(x) = \exp\left(-\frac{\|x^2\|}{2\sigma^2}\right)/2\pi\sigma^2$ is Gaussian weighting function.

4). Stereo Rendering: generate the right image or frame of a fictitious Stereo pair using the monocular query and the smoothed depth field followed by suitable processing of occlusions and newly-exposed areas. In order to generate an estimate of the right image or frame $\widehat{Q_R}$ from the monocular query Q, we need to compute a disparity δ from the estimated depth$\hat{d}$.Assuming that the fictitious frame pair (Q,QR) was captured by parallel cameras with baseline B and focal length f , the disparity is simply $\delta[x,y] = Bf/\hat{d}(x)$ where $X = [X,Y]^T$ We forward-project the 2D query Q to produce the right frame:

$$QR[X + \delta[x,y], y] = Q[x,y]$$

For this purpose we apply simple in painting using in paint _ nans or warping function from *Mat lab* Central.Figure.2 shows the input frame and its corresponding 3D view.

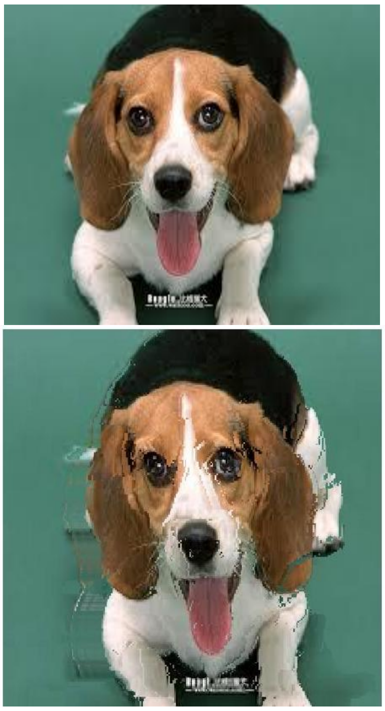

Fig.2.input frame and its 3D view

V. CONCLUSION

The paper have proposed a new method aimed at 2D-to-3D video conversion that are based on the radically different approach of learning from examples .The method is based on globally estimating the entire depth field of a query directly from a repository of frame +depth pairs using nearest neighbor-based regression. While the local method was outperformed by other algorithms ,it is extremely fast. However, our global method performed better than the state-of-the-art algorithms in terms of cumulative performance across many videos, and has done so at a fraction of CPU time. The algorithms result in a comfortable 3D experience . but are not completely void of distortions. Clearly, there is room for improvement in the future. With the continuously increasing amount of 3D data on-line and with the rapidly growing computing power in the cloud, the proposed framework seems a promising alternative to operator-assisted 2D-to-3D image and video conversion.

REFERENCES

[1] Janusz Conrad, Fellow, IEEE, Meng Wang, Prakash Ishwar, Senior Member, IEEE , Chen Wu, and Debargha Mukherjee, Learning-Based, Automatic 2D-to-3D Image and Video Conversion, IEEE transactions on Image processing, vol. 22, no. 9, september2013

[2] L .An got, W.-J. Huang , and K.-C. Liu, “ 2D to 3D video and image conversion technique based on bilateral filter,” *Proc. SPIE*, vol. 7526, p. 75260D, Feb. 2010.

[3] J. Conrad , 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.

[4] A. Saxena , M. Sun, and A. Ng, “Make3D: Learning 3D scene structure from a single still image,” *IEEE Trans. Pattern Anal. Mach.* Intel., vol. 31, no. 5, pp. 824–840, May 2009.

[5] F. Durand and J. Dorsey, “Fast bilateral filtering for the display of high-dynamic-range images,” *ACM Trans. Graph.*, vol. 21, pp. 257–266, Jul. 2002..

[6] 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.

[7] 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.

[8] 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.

[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] 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.

[11] M. Subbarao and G. Surya, “Depth from defocus: A spatial domain Approach ,” *Int. J. Comput. Vis.*, vol. 13, no. 3, pp. 271–294, 1994.

[12] 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.

[13] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection in Proc. IEEE Conf. Comput. Vis. Pattern *Recognit.*,, Jun. 2005, pp. 886–893