ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)
Vol. 2, Issue 2, Ver. 2 (April - June 2014)
International Journal of Advanced Research in Computer Science & Technology (IJARCST 2014)

A Global Nearest-Neighbor Depth Learning Based, Automatic Query to Stereo Image Conversion

[I]Priyanka T.V, [II]Pradeep Kumar N.S
[I]M.Tech (DCN), E&C Dept., S.E.A.C.E.T., Bangalore, India
[II]Assistant Professor, E&C Dept., S.E.A.C.E.T., Bangalore, India

Abstract
In spite of important growth in the availability of the stereo content, accessibility of stereo content is still insignificant in comparison by that of its query counterpart. Many image and video compression methods have been proposed in order to fill the space between query and stereo counterparts. The semi-automatic method involve human operator intervention and has been successful yet that consumes much time and expensive. Another method which is automatic methods that makes use of a deterministic stereo scene model, have not achieved the same level of quality as that of semi automatic method because they rely on assumption that are often violated in practice. A new method is proposed in this paper which is based on different approach of learning the query to stereo conversion from examples. The method is globally estimating the entire depth map of a query image directly from a preservation of stereo image(image + depth pair) using a nearest neighbor regression type idea.

Keywords
3D images, Stereoscopic images, Image conversion, Nearest neighbor classification, Cross-bilateral filtering.

I. Introduction

The availability of stereo-capable hardware today, such as TVs, Blue-Ray players, and smart phones, is not yet matched by stereo content production. Although constantly growing in numbers, stereo movies are still an exception rather than a rule, and compared to query broadcasting, stereo broadcasting (mostly sports) is still minuscule. The space between stereo hardware and stereo content availability is likely to close in the future, but today there exists an urgent need to convert the existing query content to stereo content.

The typical two steps of query-to-stereo image conversion process are: depth estimation for a given query image and depth based rendering of a new image in order to form a stereopair. While the rendering step is well understood and algorithms exist that produces the good quality images, the challenge is in estimating depth from a single image (video).Therefore, throughout this paper the focus is not on the depth based rendering but on depth recovery.

There are two basic approaches to query-to-stereo conversion: one that requires a human operator's intervention and one that does not. In the former case, the so-called semiautomatic methods have been proposed where a skilled operator assigns depth to various parts of an image or video. 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).

The method used here, carry the "big data" philosophy of machine learning. In consequence, they apply to random scenes and require no manual annotation. Our data-driven approach to query-to-stereo conversion has been inspired by the recent trend to use large image databases for various computer vision tasks, such as image saliency detection and object recognition. In particular, a new class of method that is based on the radically different approach of learning the query-to-stereo conversion from examples.

The method is based on globally estimating the entire depth map of a query image directly from a repository of stereo images (image+depth pairs or stereopairs) using a nearest-neighbor regression type idea.

The paper is organized as follows. In Section II, we review the state of the art in query to stereo image conversion. In Section III, we provide details of the global approach to the conversion. In Section IV, we conclude the paper and in section References are given.

II . STATE-OF-THE-ART

There are two types of query-to-stereo image conversion methods: semi-automatic methods, that require human operator intervention, and automatic methods, that require no such help.

A. Semi-automatic methods

Method 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.

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.* [2] 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.* [1] is the application of cross-bilateral filtering to an initial depth map. Phan *et al.* [4] propose a simplified and more efficient version using scale-space random walks that they solve with the help of graph cuts. Liao *et al.* [3] further simplify operator involvement by first computing optical flow. 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 query image, which is the main step in query-to-stereo conversion, can be formulated in various ways, for example as a shape-from shading problem. 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 while a fixed camera with varying focal length permits depth-from-defocus estimation. Both are examples of the use of multiple images of the same scene

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

captured at different times or under different exposure conditions. Although such methods are similar in spirit to the method proposed here, the main difference is that while this method use images known to depict the same scene as the query image, we use all images available in a large repository and automatically select suitable ones for depth recovery.

Several electronics manufacturers have developed realtime query to stereo 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.

Recently, machine-learning-inspired techniques employing image parsing have been used to estimate the depth. Map of a single monocular image [5]. Such methods have the potential to automatically generate depth maps, but currently work only on few types of images (mostly architectural scenes) using carefully-selected training data.

In the quest to develop data-driven approaches to query to stereo conversion we have also been inspired by the recent trend to use large image databases for various computer vision tasks, such as object recognition and image saliency detection. In our first attempt, we developed a method that fuses SIFT-aligned depth maps selected from a large stereo database, however this approach proved to be computationally demanding . 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* have proposed a depth extraction method based on SIFT warping that essentially follows our initial, unnecessarily complex, approach to depth extraction.

III. Global Nearest-Neighbor Depth Learning For Query To Stereo Conversion

In this section, method that estimates the global depth map of a query image or video frame directly from a repository of stereo images (image +depth pairs or stereo-pairs) using a nearest-neighbor regression type idea.

The approach proposed here is built upon a key observation and an assumption. The key observation is that among millions of stereo images available on-line, there likely exist many whose stereo content matches that of a query input (2D) we wish to convert to stereo image (3D). We are also making an assumption that two images that are photometrically similar also have similar stereo structure (depth).we rely on the above observation and assumption to "learn" the entire depth from a repository of stereo images I and render a stereopair in the following steps:

1. **Search for representative depth fields**: find *k* stereo images in the repository I that have most similar depth to the query image, for example by performing a *k* nearest-neighbor (*k*NN) search using a metric based on photometric properties.
2. **Depth fusion**: combine the *k* representative depth fields, for example, by means of median filtering across depth fields.
3. **Depth smoothing**: process the fused depth field to remove spurious variations, while preserving depth discontinuities, for example, by means of cross-bilateral filtering.
4. **Stereo rendering**: generate the right image of a fictitious stereopair using the monocular query image and the smoothed depth field followed by suitable processing of occlusions and newly-exposed areas.

Specific details of these steps depend on the type of stereo images contained in the repository. The above steps apply directly to stereo images represented as an image+depth pair. However, in the case of stereopairs a disparity field needs to be computed first for each left/right image pair. Then, each disparity field can be converted to a depth map, e.g., under a parallel camera geometry assumption, with fusion and smoothing taking place in the space of depths. Alternatively, the fusion and smoothing can take place in the space of disparities (without converting to depth), and the final disparity used for right-image rendering. Fig. 1 shows the block diagram of our approach. The sections below provide a description of each step and some high-level mathematical detail. In these sections, Q_R is the right image which is being sought for each query image Q, while d_Q is the query depth (ground truth) needed to numerically evaluate the performance of a depth computation. Again, we assume that a stereo dataset I is available by means of laser range finding, Kinect-based capture or disparity computation. The goal is to find a depth estimate $\hat{d}$ and then a right-image estimate $\hat{Q}_R$ given a 2D query image Q and the 3D dataset I.

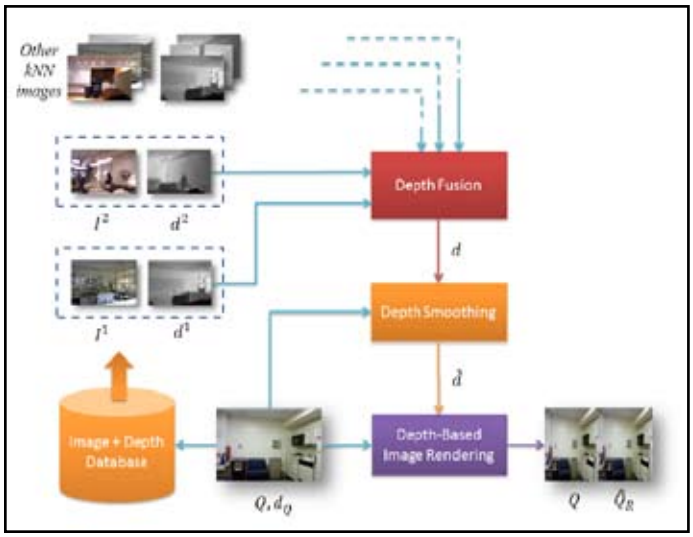

Fig. 1: Block diagram of the overall algorithm

A. kNN search

There exist two types of images in a large 3D image repository: those that are relevant for determining depth in a 2D query image, and those that are irrelevant. Images that are not photometrically similar to the 2D query need to be rejected because they are not useful for estimating depth. One method for selecting a useful subset of depth relevant images from a large repository is to select only the *k* images that are closest to the query where closeness is measured by some distance function capturing global image properties such as color, texture, edges, etc. Note that although we might miss some depth-relevant images, we are effectively limiting the number of irrelevant images that could potentially be more harmful to the query to-stereo conversion process. The selection of a smaller subset of images 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 images from a large repository is to select only the k images 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) computed from two images. Each HOG consists of 144 real values (4×4 blocks with 9 gradient direction bins) that can be efficiently computed.

We perform a search for top matches to our monocular query Q among all images $\sim Ik$, $k = 1, ..., K$ in the 3D database I. The search returns an ordered list of image+depth pairs, from the most to the least photometrically. Fig. 2 shows search results for two outdoor query images performed on the Make3D dataset #1. Although none of the four kNNs perfectly matches the corresponding 2D query, the general underlying depth is somewhat related to that expected in the query. In Fig. 2 we show search results for two indoor query images (office and dining room) performed on the NYU Kinect dataset. While some of the retained images share local 3D structures with the query image, e.g., a large table in the dining room, other images do not. Again, the general depth is somewhat related to that expected in the query. The average photometric similarity between a query and its k-th nearest neighbor usually decays with the increasing k. While for large databases, larger values of k may be appropriate, since there are many good matches, for smaller databases this may not be true. Therefore, a judicious selection of k is important.

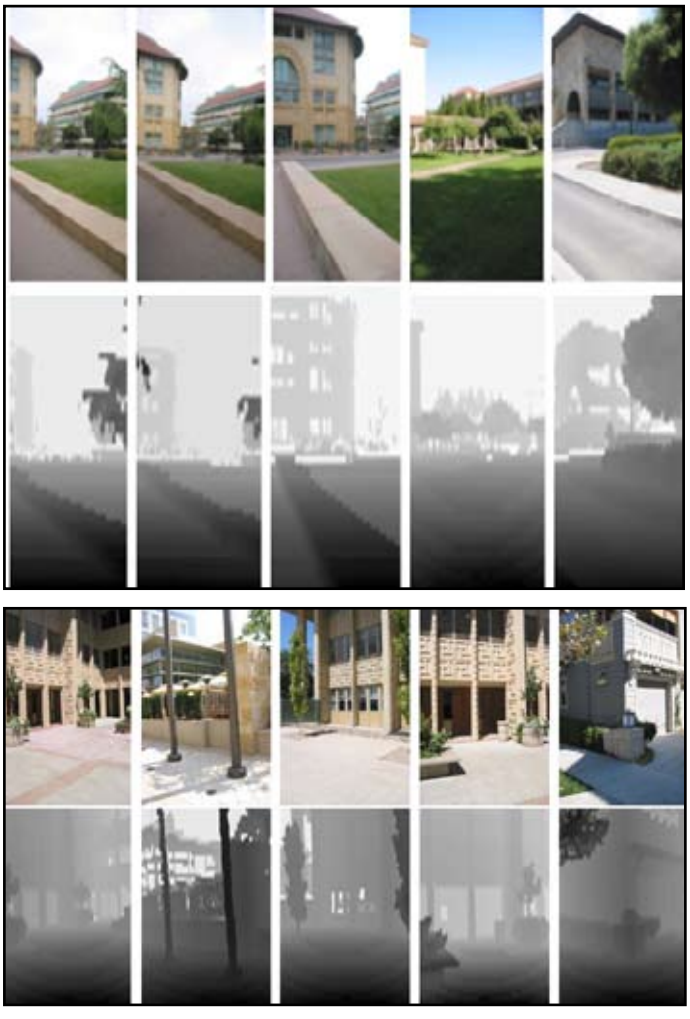

Fig. 2: RGB image and depth field of two 2D queries, and their four nearest neighbor retrieved using the Euclidean norm on the difference between histograms of gradients.

B. Depth Fusion

In general, none of the NN image+depth pairs (Ii, di), $i \in K$ match the query Q accurately. However, the location of some objects (e.g., furniture) and parts of the background (e.g., walls) is quite consistent with those in the respective query. If a similar object (e..g, building, table) appears at a similar location in several kNN images, it is likely that such an object also appears in the query, and the depth field being sought should reflect this. We compute this depth field by applying the median operator across the kNN depths at each spatial location $\mathbf{x}$ as follows:

$$d[\mathbf{x}] = \text{median}\{di[\mathbf{x}], \forall i \in K\} \quad (1)$$

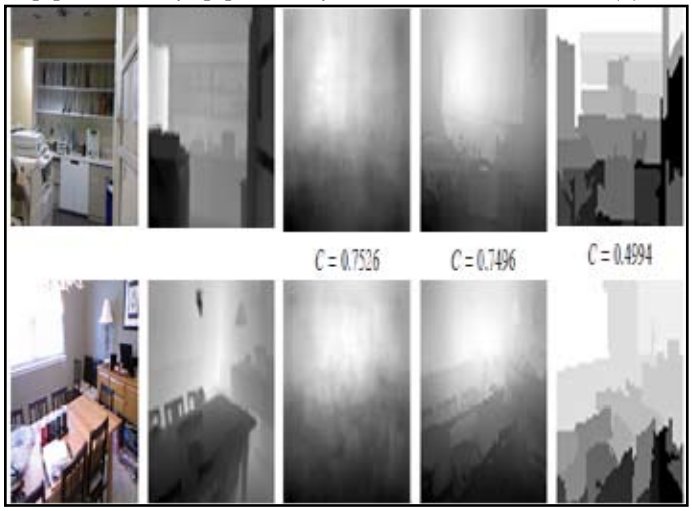

Fig. 3: Query images and depth fields: of the query, estimated depth by the global method after median-based fusion and after the same fusion and CBF, and depth computed using the Make3D algorithm.

Examples of the fused depth fields d are shown in the central column of Fig. 3 for two NYU Kinect examples. Although these depths are overly smooth, they provide a globally-correct, although coarse, assignment of distances to various areas of the scene.

C. Cross-bilateral filtering (CBF) of depth

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.

D. Stereo Rendering

In order to generate an estimate of the right image $\hat{}QR$ from the monocular query Q, we need to compute a disparity δ from the estimated depth $\hat{}d$. Assuming that the fictitious image pair $(Q, \hat{}QR)$ was captured by parallel cameras with baseline B and focal length f, the disparity is simply $\delta[x, y] = Bf/\hat{}d[\mathbf{x}]$, where $\mathbf{x} = [x, y]\ T$. We forward project the 2D query Q to produce the right image:

$$\hat{}QR[x + \delta[x, y], y] = Q[x, y] \quad (2)$$

IV. Conclusion

In this paper a new class of method aimed at query to- stereo image conversion that is based on the radically different approach of learning from examples. The method is based on globally estimating

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

the entire depth field of a query directly from a repository of image + depth pairs using nearest-neighbor based regression. We have objectively validated our algorithms' performance against state-of-the-art algorithms. Our global method performed better than the state-of-the-art algorithms in terms of cumulative performance across two datasets and two testing methods, and has done so at a fraction of CPU time. Anaglyph images produced by our algorithms result in a comfortable stereo experience but are not completely void of distortions. Clearly, there is room for improvement in the future. With the continuously increasing amount of stereo data on-line and with the rapidly growing computing power in the cloud, the proposed framework seems a promising alternative to operator-assisted query-to-stereo image and video conversion.

V. Acknowledgment

The author would like to thank all those who contributed toward making this work successful and wish to express the gratitude for the support of this project.

References

[1] *L. Agnot, W.-J. Huang, and K.-C. Liu. "A 2D to 3D video and image conversion technique based on a bilateral filter", In Proc. SPIE Three- Dimensional Image Processing and Applications, volume 7526, Feb. 2010.*

[2] *M. Guttmann, L. Wolf, and D. Cohen-Or. "Semi-automatic stereo extraction from video footage", In Proc. IEEE Int. Conf. Computer Vision, pages 136–142, Oct. 2009.*

[3] *M. Liao, J. Gao, R. Yang, and M. Gong. "Video stereolization: Combining motion analysis with user interaction", IEEE Trans. Visualization and Computer Graphics, 18(7):1079–1088, July 2012.*

[4] *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. IEEE Int. Conf. Image Processing, Sept. 2011.*

[5] *A. Saxena, M. Sun, and A. Ng. Make3D: "Learning 3D scene structure from a single still image", IEEE Trans. Pattern Anal. Machine Intell., 31(5):824 –840, May 2009.*