Buletin Teknik Elektro dan Informatika
(Bulletin of Electrical Engineering and Informatics)
Vol. 3, No. 3, September 2014, pp. 149~154
ISSN: 2089-3191

A Global Nearest-Neighbour Depth Estimation-based Automatic 2D-to-3D Image and Video Conversion

Anusha M Sidhanti, Jyothsna Madam, Mounesh V M
C/O Dr M N Sidhanti, Panchakacheri Oni
Hosayallapur, Dharwad, 580001, Karnataka, India
e-mail: anushasidhanti@gmail.com

Abstract

The proposed work is to present a new method based on the radically different approach of learning the 2D-to-3D conversion from examples. It is based on lobally estimating the entire depth map of a query image directly from a repository of 3D images (image depth pairs or stereo pairs) using a nearest-neighbour regression type idea.

Keywords: *Stereoscopic images, Nearest-neighbor classification, repository of images, 3D images*

1. Introduction

The availability of 3D-capable hardware today, such as TVs, Blu-Ray players, gaming consoles, and smartphones, is not yet matched by 3D content production [1]. Although constantly growing in numbers, 3D movies are still an exception rather than a rule, and 3D broadcasting (mostly sports) is still minuscule compared to 2D broadcasting. The gap between 3D hardware and 3D content availability is likely to close in the future, but today there exists an urgent need to convert the existing 2D content to 3D, that require human operator intervention, and automatic methods, that require no such help [2].

A typical 2D-to-3D conversion process consists of two steps: depth estimation for a given 2D image and depth based rendering of a new image in order to form a stereo pair. While the rendering step is well understood and algorithms exist that produce good quality images, the challenge is in estimating depth from a single image (video).

2. Literature Survey

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 semiautomatic 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 [3-4]. The involvement of a human operator may vary from just a few scribbles to assign depth to various locations in an image to a precise delineation of objects and subsequent depth assignment to the delineated regions. In the case of automatic methods, no operator intervention is needed. Although restricted to architectural scenes, these methods opened a new direction for 2D-to-3D conversion. There are two types of 2D-to-3D image conversion methods: semi-automatic method and automatic method.

To this effect, methods have been developed that estimate shape from shading, structure from motion or depth from defocus. Although such methods have been shown to work in some restricted scenarios they do not work well for arbitrary scenes. In an attempt to equip 3D TVs, Blu-Ray players gaming consoles. With real-time automatic 2Dto-3D conversion, consumer electronics manufacturers have developed simpler techniques that rely on various heuristic assumptions but such methods fail on more challenging scenes. Recently, machine-learning-inspired methods have been proposed to automatically estimate the depth map of a single monocular image by applying image parsing.

Received March 9, 2014; Revised May 28, 2014; Accepted June 12, 2014

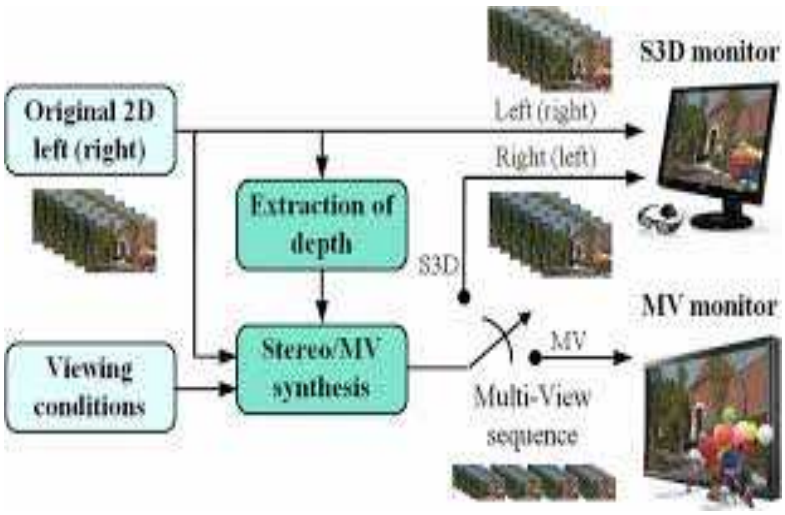

Figure 1. A new direction for 2D-to-3D conversion

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 labour-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 that 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 role of an operator is to correct errors in the automatically computed depth of moving objects and assign depth in undefined areas.

Automatic methods

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 However, this problem is severely under-constrained; quality depth estimates can be found only for special cases. Other methods, often called multiview stereo, attempt to recover depth 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 far away.

Semi atomic method 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.

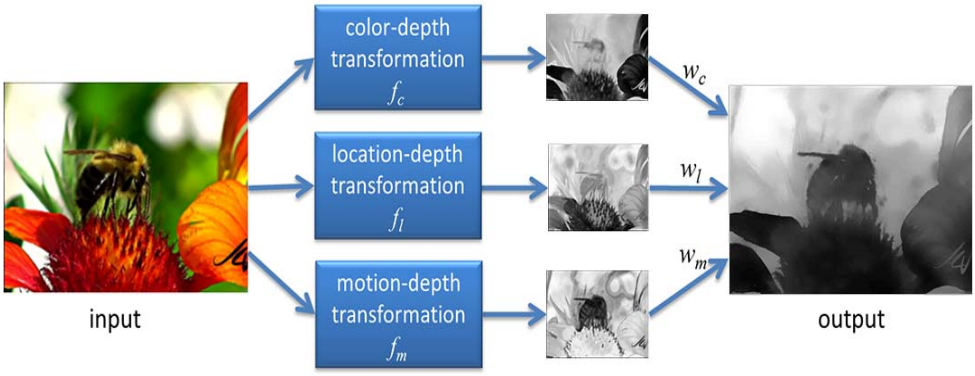

Figure 2. Example of depth estimation from color, spatial location and motion. Black indicates smallest depth

In the quest to develop data-driven approaches to 2D-to-3D 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. 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.

In this algorithm video conversion methods have been proposed. It is successful but also time-consuming and costly. Automatic methods, that typically make use of a *deterministic* 3D scene model, yet achieved the same level of quality for they rely on assumptions that are often violated in practice. Thus we propose a new method which is based on the approach of *learning* the 2D-to-3D conversion from examples. The point transformation can be learned off-line and applied basically in real time – the same transformation is applied to images 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 colour, spatial position, and motion at each pixel To address this limitation, in this section we develop a second method that estimates the *global* depth map of a query image or video frame directly from a repository of 3D images.

3. The Proposed System

2D-to-3D conversion by learning a local point transformation

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. We are also making an assumption that two images that are photometrically 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.

To “learn” the entire depth from a repository of 3D images and render a stereo pair in the following steps:

Search for representative depth fields: Find 3D images in the repository I that have most similar depth to the query image, for example by performing a *k* nearest-neighbour **(*k*NN)** search using a metric based on photometric properties.
Depth fusion: Combine the *k* representative depth fields, for example, by means of median filtering across depth fields.

Depth smoothing: Process the fused depth field to remove spurious variations, while preserving depth discontinuities, for example, by means of cross-bilateral filtering.

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.

The above steps apply directly to 3D 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.

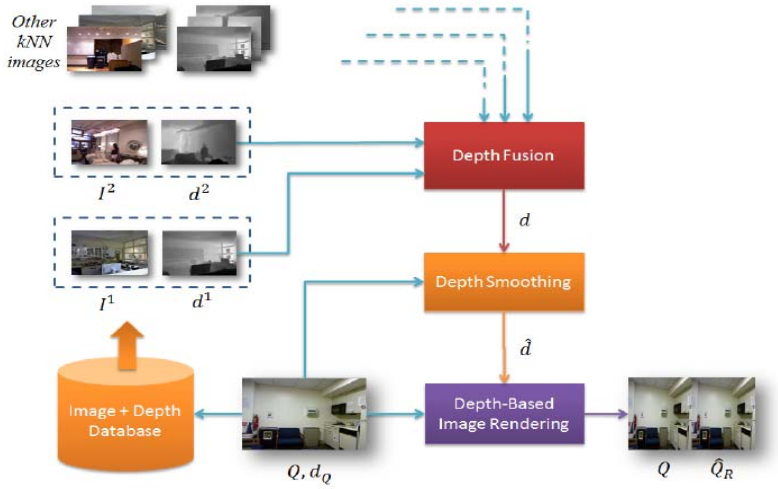

Figure 3. Block diagram of the overall algorithm

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 (as per our assumption). 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 2D-to-3D 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) [3] computed from two images. Each HOG consists of 144 real values (4×4 blocks with 9 gradient direction bins) that can be efficiently computed.

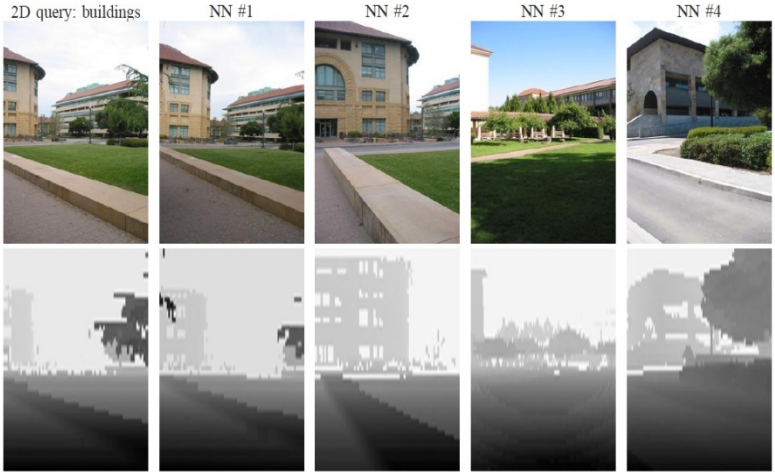

Figure 4. kNN search process

Depth fusion

If a similar object (e..g, building, table) appears at a similar location in several *k*NN 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 *k*NN depths at each spatial location **x** as follows:

$$d[\mathbf{x}] = \text{median}\ \{di[\mathbf{x}],\ i\ K\}.$$

Although these depths are overly smooth, they provide a globally-correct, although coarse, assignment of distances to various areas of the scene.

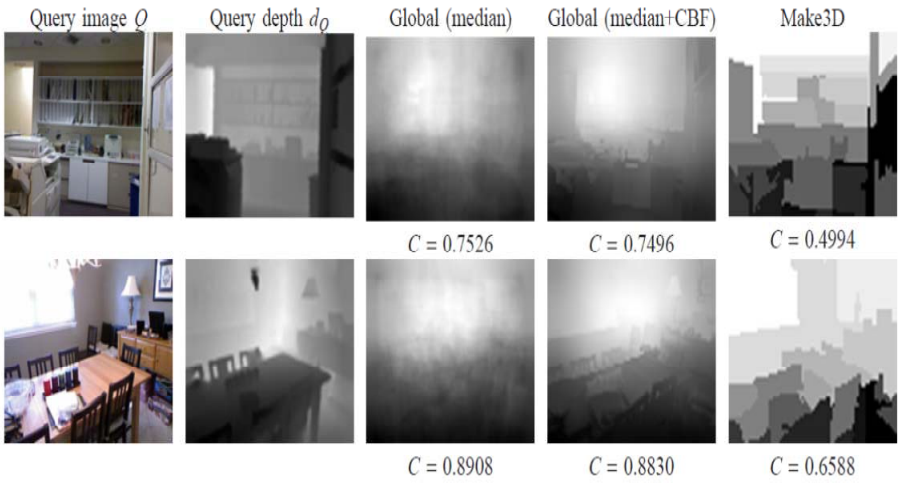

Figure 5. Depth fusion process

Figure below shows the 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. Normalized depth cross-covariance are included under each estimated depth field.

Cross-bilateral filtering (CBF) of depth

While the median-based fusion helps make depth more consistent globally, the fused depth is overly smooth and locally inconsistent with the query image due to edge misalignment

between the depth fields of the *k*NNs and the query image. 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, 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 [4]. 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 *d* using the query image *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 image *Q* and local noise/granularity suppression in the fused depth.

4. Conclusion

We have proposed a new class of methods aimed at 2Dto-3D image conversion that are based on the radically different approach of learning from examples. One method that we proposed is based on learning a point mapping from local image attributes to scene-depth. The other method is based on globally estimating 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. While the local method was Outperformed by other algorithms, it is extremely fast as it is, basically, based on table look-up. However, 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 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.

References

[1] L Agnot, WJ Huang, KC Liu. *A 2D to 3D video and image conversion technique based on a bilateral filter*. SPIE Three-Dimensional Image Processing and Applications. Vol. 7526. 2010.
[2] T Brox, A Bruhn, N Papenberg, J Weickert. *High accuracy optical flow estimation based on a theory for warping*. Proc European Conf. Computer Vision. 2011: 25–36.
[3] N Dalal, B Triggs. *Histograms of oriented gradients for human detection*. Proc. IEEE Conf. Computer Vision Pattern Recognition. 2010: 886–893.
[4] F Durand, J Dorsey. Fast bilateral filtering for the display of high-dynamic-range images. *ACM Trans. Graph.* 2012; 21: 257–266.