J. Vis. Commun. Image R. 43 (2017) 127–137

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

# Key-frame-based depth propagation for semi-automatic stereoscopic video conversion ☆

Guo-Shiang Lin [a], Jian-Fa Huang [b], Wen-Nung Lie [b,*]

[a] Dept. of Computer Science and Information Engineering, Da-Yeh University, Taiwan, ROC
[b] Dept. of Electrical Engineering, National Chung Cheng University, Taiwan, ROC

ARTICLE INFO

*Article history:*
Received 8 August 2016
Revised 29 October 2016
Accepted 8 December 2016
Available online 24 December 2016

*Keywords:*
2D-to-3D stereo video conversion
Depth propagation
Key-frame
Error correction
Depth image based rendering (DIBR)

ABSTRACT

In this paper, we propose a key-frame-based bi-directional depth propagation algorithm for semi-automatic 2D-to-3D stereoscopic video conversion. First, key-frames are identified from each video shot based on color motion-compensation errors to prevent high-motion content between any pair of consecutive key frames. Depths for key-frames are manually assigned or rendered by popular computer tools, and then bi-directionally propagated to non-key-frames there between. Our depth propagation algorithm is featured of a multi-pass error correcting procedure for each frame to prevent depth artifacts from being further propagated to adjacent frames. Our proposed algorithm is advantageous in solving the background occlusion/dis-occlusion problem that degrades the performances of traditional depth propagation algorithms. Experimental results show that our scheme is capable of achieving better results against three prior algorithms in view of the qualities of the estimated depth map (e.g., dis-occluded background and object boundaries) and the synthesized stereo views.



## 1. Introduction

3D TV has become one selling point of high-end TVs since 2010 [18,25]. However, according to the reports, consumers use TV's 3D function at only about 1% of time, due to the lack of 3D video content. Although film companies have been using 3D cameras to shoot three-dimensional movies, the expensive equipment and production process prevent mass production of 3D content. For 3D content providers, it might be the most efficient way to produce 3D versions from existing 2D video database via Stereoscopic Video Conversion (SVC), also called 2D-to-3D conversion, techniques [3,11,14].

In general, SVC techniques lay their emphasis on depth estimation for each frame. A combination of color plus depth information [15,31] via the Depth-Image-Based Rendering (DIBR) technique [1,3] can lead to novel view synthesis and thus stereo display. From the viewpoint of human's aid, SVC methods can be categorized into three classes: manual, semi-automatic, and fully automatic [3,14]. Manual assignment of depths for each individual frame can produce videos of good 3D effect, but is considerably time-consuming (normally, spending several minutes for one frame). The automatic methods calculate monocular or binocular image clues from such as motion, linear perspective, atmospheric perspective, texture gradient, and relative size, to estimate 3D geometries about the scene [3–5]. Though automatic SVC methods can convert existing 2D videos into 3D content in a more efficient manner, they suffer from inaccurate depth estimation and less adaptation to varying video content, which may degrade observers' 3D perception. This motivates us to develop a semi-automatic SVC (SA-SVC) technique to generate stereoscopic videos.

Considering high-quality 3D video for public use (e.g., DVD rent store or VoD system), SA-SVC techniques [2,7–10,16–17] are devised to overcome drawbacks of manual or automatic methods. In this type of processing, a small set of key frames is first chosen for manual assignment of depth, where some computer tools (e.g., Photoshop and PhotoImapct, etc.) or algorithms (e.g., Lazy snapping [13]) were used to facilitate user's operations, which is beyond the discussion of this paper. After manual depth assignment, key frames and their depth maps are then used to estimate the depth maps of other non-key frames for reducing the overall production cost. This will of no doubt form the most critical part of an SA-SVC system.

Traditional SA-SVC techniques can be classified into two classes: learning-based [16–17] and propagation-based [2,7–10]. In [16,17], machine learning algorithms were developed to learn a classifier for the relationship between depths and image features

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding author.
*E-mail addresses:* khlin@mail.dyu.edu.tw (G.-S. Lin), ieewnl@ccu.edu.tw (W.-N. Lie).

http://dx.doi.org/10.1016/j.jvcir.2016.12.002
1047-3203/

of key-frame pixels. The trained classifier is then used to classify depths for pixels of non-key frames. This kind of algorithms is however dis-advantageous in producing correct depths at discontinuities. The methods proposed in [16,17] simply use local image features, spatial positions, and color values to estimate depths of non-key frames. Though simple, it seems difficult to obtain a well-trained classifier that is capable of accurately reflecting or predicting depths of pixels that are not in the training set.

Being one of the propagation-based methods [2,7–10], Wu et al. [7] first tracks contours of foreground objects between successive frames and then performs depth interpolation based on the tracked contours. Their method cannot process videos with camera motion since the background depths are assumed static or zero (i.e., infinitely far). In [6], Zhang et al. proposed an interactive system to achieve stereoscopic video conversion, where the foreground depth map of each key frame is initially generated based on several mono-view depth cues and then refined by a human-assisted interactive procedure. Foreground depth information in the key frames is then propagated to non-key frames according to motion information. Finally, contour of the propagated foreground boundary is refined by using an adaptive level set method. Varekamp and Barenbrug [8] proposed to propagate depth information of the start key-frame to subsequent non-key frames via motion compensation (MC) [23] on a block-by-block basis, thus is in no need of prior object segmentation. Their method first estimates an initial depth map at time $t$ by bilateral filtering, which is then followed by depth motion estimation (ME) and compensation (MC) between time $t$ and $t$-1. Depth propagation in a single forward direction [8] however causes degraded qualities in depths for non-key frames with large temporal distances. To improve the performance of [8], motion estimation and bilateral filtering [20] were modified [9,10,28]. and bi-directional depth propagations from two closet surrounding key frames was performed in [9,10]. Their methods are however limited to a linear combination of the forward and backward depth propagation results, where the combination coefficients are determined by the temporal distances from each bounding key frame. This kind of simple linear combination nevertheless cannot resolve the propagation artifacts due to textureless regions, large object motions, camera motion, or occlusions/dis-occlusions around object contours.

Essentially, there exist two manners of depth propagation, one is key-frame-centralized (KFC) and the other is ripple-like (RIP), as illustrated in Fig. 1, where a video segment $S$ with two bounding key frames is given. Depth maps $\{D_0 \sim D_{|S|-1}\}$ ($|S|$: the number of frames in $S$) corresponding to the color frames of the video segment are to be derived. Normally, $D_0$ and $D_{|S|-1}$ are manually drawn/assigned and $D_t$, $0 < t < |S| - 1$, is estimated via direct/indirect propagation from $D_0$ and $D_{|S|-1}$ of two key frames in both forward and backward directions. For RIP schemes ([8,10,27,30,32] belong to this kind), $D_t$ is estimated from $D_{t-1}$ and $D_{t+1}$ in the forward and backward direction, respectively. For KFC schemes ([9,26,29] belong to this kind), $D_t$ is obtained directly via propagations from $D_0$ and $D_{|S|-1}$ of the two key frames. The KFC schemes are advantageous of reducing depth error propagations from other non-key frames, but suffering from a shorter distance allowed between the key and non-key frames due to limited motion search ranges and inaccurate matching. KFC schemes thus subsequently increase the number of key frames that need manual assignment of depths. To have a compromise between the KFC and RIP schemes, our prior work [32] adopts a strategy that the depth of each frame should be incurred a two-pass bi-directional depth propagation process to ensure less error propagation and better depth quality. Here, in this paper, we extend the prior work [32] to perform a multi-pass error correction process (mainly for image blocks of dis-occlusion and object boundary, see Section 4.3) for each depth frame before it is propagated to the next frame in the specified directions (forward and backward), thus alleviating possible error propagations. Though Wang et al. [29] revealed a similar concept of correcting MVs before depth propagation, their MVs are however estimated based on a cost function combining the color and depth matching errors in uni-directional referencing (traditionally, backwards) and did not consider bi-directional estimation to resolve image blocks of dis-occlusion and object boundary.

The remainder of this paper is organized as follows. Section 2 describes the system architecture of our scheme. Sections 3 and 4 elaborate the key-frame selection and multi-pass depth propagation steps of our SA-SVC scheme, respectively. In Section 5, experiment results are given and finally Section 6 draws some conclusions and future work.

## 2. System architecture

To achieve SA-SVC, each video shot is partitioned into segments, separated by intermediate key frames, with the first and last frames being considered as the default bounding key-frames (see Fig. 2). For example, an image sequence composed of a moving dancer on a stage is considered as a single shot, but partitioned into several segments, depending on the motion activity of the dancer. We restrict a segment to be a series of frames that lead to acceptable depth propagation. Since our depth-propagation algorithm is motion-compensation-based (see later descriptions), the selection of intermediate key frames is highly related to object motions.

Fig. 3 shows the system processing flow. First, key frames are determined based on a fixed-interval strategy or result of color frame analysis. The identified key frames are then manually assigned with depth maps, by which depth maps of in-between non-key frames are estimated via bi-directional depth propagation (as in Fig. 1). All the produced depth maps, together with the input color texture frames, can then be used for stereo view synthesis (e.g., via DIBR).

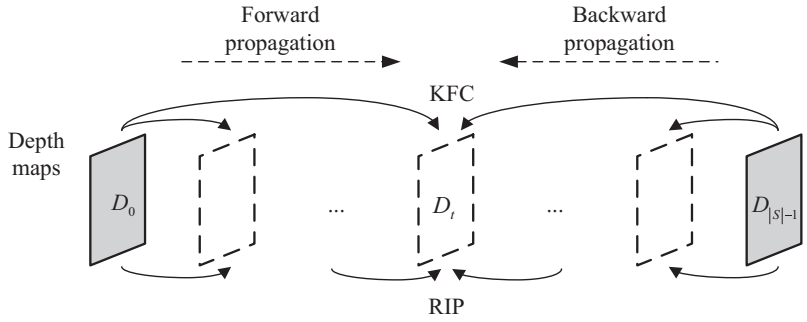


**Fig. 1.** Illustration of bi-directional depth propagation for frames in video segment $\{D_0 \sim D_{|S|-1}\}$.