# Toward Naturalistic 2D-to-3D Conversion

Weicheng Huang, Xun Cao, *Member, IEEE*, Ke Lu, Qionghai Dai, *Senior Member, IEEE*, and Alan Conrad Bovik, *Fellow, IEEE*

***Abstract*— Natural scene statistics (NSSs) models have been developed that make it possible to impose useful perceptually relevant priors on the luminance, colors, and depth maps of natural scenes. We show that these models can be used to develop 3D content creation algorithms that can convert monocular 2D videos into statistically natural 3D-viewable videos. First, accurate depth information on key frames is obtained via human annotation. Then, both forward and backward motion vectors are estimated and compared to decide the initial depth values, and a compensation process is applied to further improve the depth initialization. Then, the luminance/chrominance and initial depth map are decomposed by a Gabor filter bank. Each subband of depth is modeled to produce a NSS prior term. The statistical color–depth priors are combined with the spatial smoothness constraint in the depth propagation target function as a prior regularizing term. The final depth map associated with each frame of the input 2D video is optimized by minimizing the target function over all subbands. In the end, stereoscopic frames are rendered from the color frames and their associated depth maps. We evaluated the quality of the generated 3D videos using both subjective and objective quality assessment methods. The experimental results obtained on various sequences show that the presented method outperforms several state-of-the-art 2D-to-3D conversion methods.**

***Index Terms*— 2D-to-3D conversion, depth propagation, natural scene statistics, Bayesian inference.**

## I. Introduction

THREE DIMENSIONAL (3D) video has become quite popular in recent years. Yet, the proliferation of 3D capture and display devices has not been matched by a corresponding degree of availability of quality 3D video content. Towards helping to overcome this 3D content shortage, a new 3D content creation technology, 2D-to-3D conversion, is being developed to convert existing 2D videos into 3D videos [1], [2].

2D-to-3D video conversion methods can be divided into two categories, depending on whether human-computer interactions are involved in the conversion process: fully-automatic methods and semi-automatic methods [2]. Current fully-automatic methods are generally only able to deliver a limited 3D effect. However, semi-automatic methods have made it possible to balance 3D content quality with production cost, and has been demonstrated to enable the conversion of popular old films – such as the *Star Wars* series, *Titanic*, and so on into successful cinematic 3D presentations [3]. The general approach to semi-automatic 2D-to-3D conversion is to manually or semi-manually create high quality depth maps at strategically chosen key frames or parts of frames, then propagate depth information from the key frames to non-key frames to initiate depth calculations at non-key frames (see Fig. 1 as an illustration). The highest cost arises during the process of assigning depths to key frames, whereas the 3D quality of the final production largely depends on the accuracy of the key frame depth maps, the key frame separations, and the depth propagation method. Smaller intervals and higher key depth accuracy lay a better foundation for subsequent depth propagation, leading to improved stereo quality. Unfortunately, these increase the cost as well.

Developing depth propagation methods that effectively control depth errors can make it possible to relax the key frame interval constraints, while also significantly improving the final quality. Of course, the additional algorithm complexity of automation is negligible as compared with the reduction in human-computer interaction. This is the main reason why depth propagation plays such a critical role in 2D-to-3D video conversion.

Recently, statistical models of natural scenes have proven to provide useful constraints on many image processing and computer vision problems, including image compression [4], image and video quality prediction [5], image denoising [6] and stereo matching [7], [8]. They provide powerful statistical priors that can force ill-posed visual problems towards stable, naturalistic solutions. For example, the univariate distributions of band-pass luminance images (wavelet coefficients) are well-modeled as obeying a generalized Gaussian distribution:

$$P(c) = \frac{e^{-|c/s|^p}}{Z(s,p)} \tag{1}$$

where $Z(s, p)$ is a normalizing constant that forces the integral of $P(c)$ to be 1, while the parameters $p$, $s$ control the shape and spread of the distribution, respectively. Liu *et al.* [7] also showed that the conditional magnitudes of luminance and depth are mutually dependent, *i.e.* regions exhibiting larger luminance variations often have larger depth variations and vice versa.

Manuscript received October 15, 2013; revised May 30, 2014; accepted December 17, 2014. Date of publication December 23, 2014; date of current version January 9, 2015. This work was supported in part by the National Science Foundation of China under Project 61371166 and Project 61422107, in part by the Importation and Development of High-Caliber Talents Project through the Beijing Municipal Institutions under Grant IDHT20130225, in part by the National Natural Science Foundation of China under Grant 61103130 and Grant 61271435, and in part by the National Program on Key Basic Research Project (973 Program) under Grant 2010CB731804-1. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Charles Boncelet.

W. Huang and K. Lu are with the College of Engineering and Information Technology, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: luk@ucas.ac.cn).

X. Cao is with the School of Electronic Science and Engineering, Nanjing University, Nanjing 210093, China (e-mail: caoxun@nju.edu.cn).

Q. Dai is with the Department of Automation, Tsinghua University, Beijing 100084, China (e-mail: qhdai@tsinghua.edu.cn).

A. C. Bovik is with the Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, TX 78712 USA (e-mail: bovik@ece.utexas.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2014.2385474

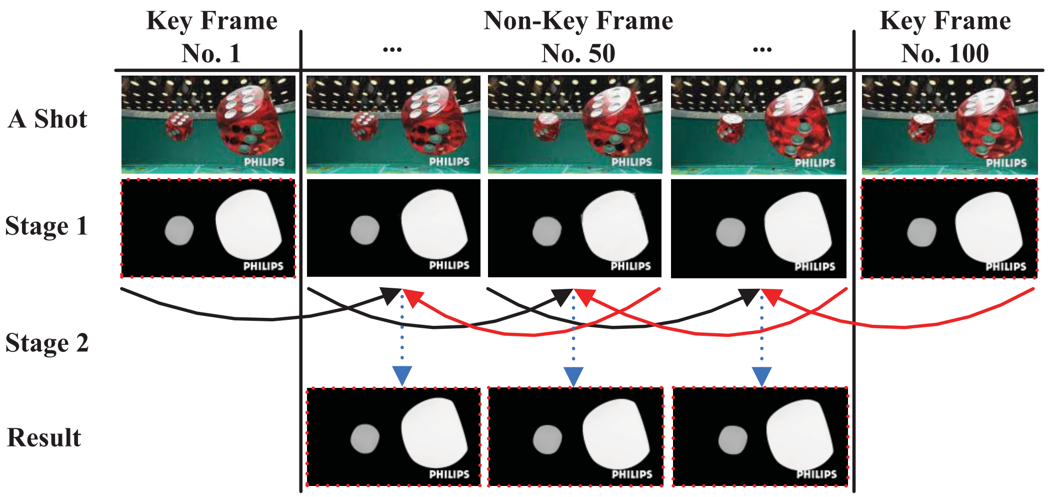


Fig. 1. In stage 1, the initial depths of non-key frames are propagated from previous and next frames. The solid arrows denote bi-direction motion estimation and compensation. In stage 2, the dashed arrows denote refinement of the initial depths.

Towards producing high-quality and cost-effective 3D videos, we have developed a semi-automatic 2D-to-3D conversion method that utilizes statistical depth priors to stabilize computation and provide naturalistic solutions. The overall depth computation framework is cast in a Bayesian framework. Greatly extending our recent preliminary work on NSS-guided depth propagation using only luminance information [9], we further our contribution in this direction in three ways:

- We explore the statistical relationship between luminance/chrominance and depth in natural images [10]. Chromatic information can be used to reveal perceptually relevant statistical relationships between scene texture and geometry, leading to higher converted 3D video quality.
- We also exploit spatial correlations that exist in the depth map itself. Natural scene depth maps tend to be largely smooth and coherent, exhibiting sparse, connected discontinuities.
- Going beyond the simple Mean-Square-Error (MSE) test in [9], solid validation is conducted via both subjective and objective quality assessment demonstrating the effectiveness of our new statistics-driven model method over several state-of-the-art 2D-to-3D conversion schemes.

We proceed as follows: Section II reviews previous 2D-to-3D conversion systems; Section III introduces the entire pipeline of 3D video generation from monocular 2D video; our proposed method of statistical and structural modeling of depth/range and luminance/chrominance is explained in Section IV; our experimental design of 3D quality evaluation along with the performance results are presented in Section V. Finally, we conclude Section VI with thoughts towards future work.

## II. Related Work

Current 3D content generation methods include stereoscopic photography and 3D scene modeling using commercial software. A third method, 2D-to-3D conversion, continues to evolve and gain acceptance because of its efficiency, flexibility and steerability.

Fully-automatic 2D-to-3D conversion methods can be used in situations where human interaction with the video is not possible, for example, real-time conversion of live TV broadcasts. Estimating missing depth data from just one piece of 2D video is a difficult ill-posed problem. Monocular depth cues can be utilized to estimate 3D information [2], for instance, linear perspective, occlusion, texture gradient, and motion parallax [11], [12]. Another strategy to automatically generate depth values is by training on a large domain-specific dataset, using known statistical relationships between the texture and depth information. Saxena *et al.* devised a supervised learning strategy for predicting depth from images [13]. The use of semantic labels in the learning process was shown to produce better depth estimates in [14]. Temporal coherence features were used in [15] to achieve promising 3D conversion results using a Kinect collected "color + depth" database. However, fully-automatic methods are still in the early stages of development.

Aiming at finer 3D quality, semi-automatic 2D-to-3D conversion methods exploit user interactions to provide initial depth estimates at key frames. After that, the main task is propagating the annotated depths over time. Recent semi-automatic methods include Varekamp *et al.* [16] who propose a scheme which generates non-key frame depths via bilateral filtering, then refines them using a block-based motion compensation algorithm. Wu *et al.* [17] propose a depth propagation method using bi-direction optical flow and the Mean Shift algorithm for foreground object extraction and non-key frame depth tracking. Cao *et al.* [18] proposed a semi-automatic conversion method employing a multiple object segmentation algorithm to create disparity maps for key frames and shifted bilateral filtering to propagate disparities to non-key frames. Recently, a real-time optimization accelerated by a GPU was demonstrated which enables users to monitor the 3D effect of just-draw scribbles on-the-fly [19].

While progress in 2D-to-3D conversion holds considerable promise for 3D content generation, previous work has not

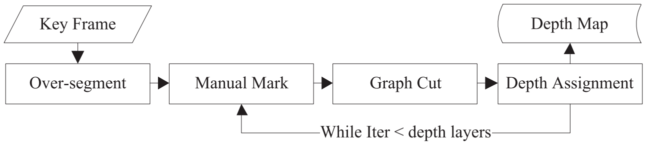


Fig. 2. Pipeline of key-frame depth map assignment.

addressed the interplay between natural depth, luminance, and color statistics and how they relate to the creation of depth propagated 3D videos. These are the contributions explained in this paper.

## III. Method Overview

The basic data unit when conducting semi-automatic 2D-to-3D video conversion is the shot, or series of consecutive pictures taken continuously by a single camera and representing connected actions in time and space. The first and last frame in a video shot are called key frames, while frames between them are called non-key frames. Many well developed shot detection algorithms can be used to partition a video into shots such as methods based on color histogram differences [20] or edge changes [21].

An example is shown in Fig. 1. The top row is an image sequence in a pre-segmented shot. In Stage 1, we assign depth only to key frames, and estimate the depth of non-key frames. In Stage 2, only non-key frame depths are refined, yielding the final depth sequence. The main contribution occurs in the center of Fig. 1: initial non-key frame depth estimation, and depth refinement under natural scene statistic prior constraints.

### A. Initial Depth Estimation

Assume that accurate depth maps have been obtained for every key frame. This can be accomplished via an involved human-computer interaction [18]. First, an over-segmentation map and a connection map of the over-segmented areas are built. Some areas are assigned as foreground and background by user's scribbles, while others are assigned using graph-cut optimization [22]. Finally, user can assign depth to each object layer with the aid of preset depth models. Fig. 2 shows the pipeline of human annotation for depth assignment on key frames.

Before the next refinement stage, initial depth estimates for non-key frames must be found. To accomplish this we deploy the method proposed in [18] and [23]. First, both forward and backward motion vectors are estimated between previous and current frames. Then direct copy and shifted bilateral filtering is applied to successful and failed matched pixels, respectively. After that, depths are propagated forward and backward across all non-key frames. Merging is then carried out at each non-key frame using a weighting strategy that decreases with propagation distance.

In the proposed method, the temporal consistency is partially fulfilled through the smoothed motion vector. In one video shot, since the depth values are propagated from the same key frames, the temporal consistence of the motion vector leads to smoothness in the depth domain.

This stage is illustrated in Fig. 3.

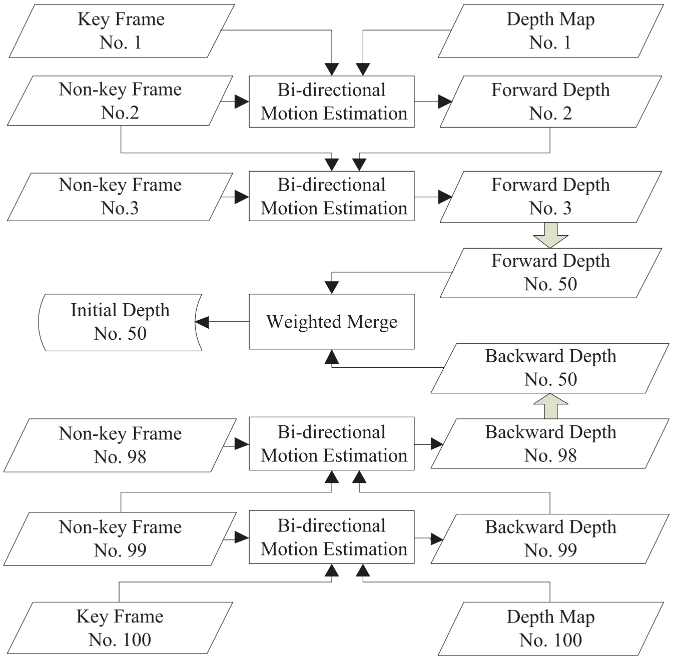


Fig. 3. Pipeline of initial depth estimation.

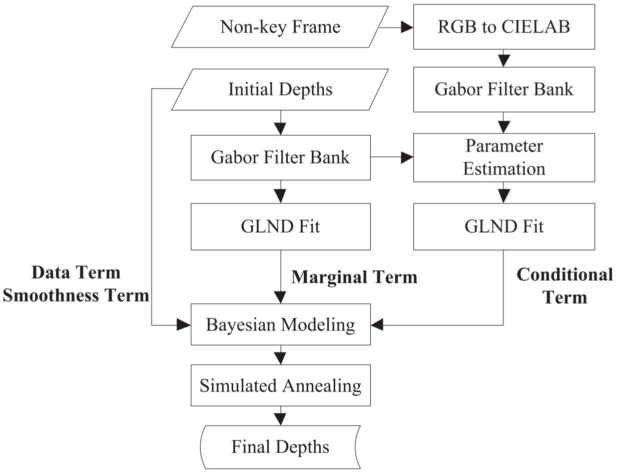


Fig. 4. Pipeline of depth refinement.

### B. Depth Refinement

The depth refinement stage is designed to produce a high-quality, naturalistic depth sequence from the raw initial depth estimates. Our method for doing this uses a Bayesian natural scenes statistics based prior model. As shown in Fig. 4, the depth refinement stage includes the following steps, where GLND Fit is the best least-squares fit of a generalized log-normal distribution (GLND) to the empirical distributions of the Gabor filter magnitude responses.

Given a non-key frame, a depth refinement algorithm processes the initial depth map estimated from the previous stage into a suitable depth estimate at the current non-key frame. The basic idea is to minimize an energy functional which merges differential depth cues between the current color image and the previous depth map within an optimization framework [24]. A Bayesian inference algorithm is formulated that incorporates a likelihood (conditional distribution) and a prior (marginal distribution) of the natural scene statistics (NSS) model within the energy function to be minimized, thus forcing the solution to be consistent with the observed statistical relationships that occur between luminance, chrominance, and depth in natural scenes.

First, we convert the non-key frames to the perceptually uniform CIELAB color space. CIELAB was designed so that color changes corresponding to a given distance within the color space will yield a similar perceived amount of change [10]. Then, both luminance/chrominance and the initial depth map are decomposed by a Gabor filter bank defined by two radial center frequencies and four orientations. A GLND natural scene statistics model of the Gabor magnitude responses is applied to the normalized histogram of each sub-band by a least-squares fitting progress. This is incorporated as a prior that is appended to the overall likelihood function to measure the departure from "naturalness" of the Gabor filter responses.

The data term and smoothness term are likewise introduced into the energy function in a manner similar to stereo matching algorithms. The data term measures how well the refined depths fit the initial depth maps, while the smoothness term controls the penalty on depth variations.

Finally, the energy function is optimized using the Simulated Annealing algorithm.

## IV. Bayesian Inference Using NSS

### A. Gabor Filter Bank

The statistical analysis is conducted on the Gabor magnitude responses to luminance, chrominance, and depth. We adopted the same Gabor filter bank as used in [8]. A real Gabor function can be written

$$G_{\lambda,\theta,\psi,\sigma,\gamma}(x,y) = \exp\left(-\frac{u^2+\gamma^2 v^2}{2\sigma^2}\right)\cos\left(2\pi\frac{u}{\lambda}+\psi\right) \quad (2)$$

where $u = x\cos\theta + y\sin\theta$ and $v = -x\sin\theta + y\cos\theta$. The standard deviation $\sigma$ and aspect ratio $\gamma$ determine the size and eccentricity of the elliptical Gaussian envelope, respectively. $\lambda$ is the carrier wavelength, $2\pi/\lambda$ is the spatial frequency of the sinusoidal carrier, and $\theta$ is the filter orientation. Finally, the phase offset $\psi$ specifies the symmetry of the Gabor function: the function is even symmetric when $\psi = 0$ or $\psi = \pi$, and odd symmetric when $\psi = -\pi/2$ or $\psi = \pi/2$.

Considering that mid-frequencies provide the best performance when linear-fitting the depth response and the color distribution parameters, two center frequencies, 2.22 and 3.61 (cycles/degree), are used, with four different sinusoidal grating orientations for each spatial frequency: horizontal (0), diagonal-45 ($\pi/4$), vertical ($\pi/2$), and diagonal-135 ($3\pi/4$) are used in [8]. The aspect ratio, $\gamma$, is chosen to be 1.0 [25]. The spatial frequency bandwidth of each sub-band is 0.7 (octave), and neighboring filters intersect at half-power point, *i.e.* 3-dB point [26]–[29]. The half-response spatial frequency $b$ and the ratio $\sigma/\lambda$ are related as follows:

$$\frac{\sigma}{\lambda} = \frac{1}{\pi}\sqrt{\frac{\ln 2}{2}}\cdot\frac{2^b+1}{2^b-1} \quad (3)$$

Given a spatial frequency and an orientation, the Gabor magnitude response for each pixel is

$$\tilde{I}(i,j) = \sqrt{\tilde{I}_o(i,j)^2 + \tilde{I}_e(i,j)^2} \quad (4)$$

where $\tilde{I}_o$ and $\tilde{I}_e$ are the results of the original image $I$ filtered by a pair of odd-symmetric and even-symmetric Gabor filters.

### B. Bayesian Modeling

Taking the initial depth estimation as a number of measurements and the true depth as unknown parameters, the refinement problem becomes an inverse problem. A maximum likelihood estimation can often lead to good solutions. To find a balance between the initial depth estimation and prior knowledge, we re-build the Bayesian formulation to integrate the initial depth estimation, the NSS prior and the spatial smoothness constraint into a single energy functional. Each factor in the proposed energy functional can be balanced by its linear weight.

Given a non-key frame, $I$, and its initial depth map, $D_{init}$, then to estimate the final depth map, $D$, the canonical Bayesian inference formulation takes the form

$$\begin{aligned} D &= \arg\max_{D'} P\left(D' \middle| (I, D_{init})\right) \\ &= \arg\max_{D'} P\left((I, D_{init}) \middle| D'\right) P(D') \end{aligned} \quad (5)$$

where $P\left(D' | (I, D_{init})\right)$ is the posterior probability to be maximized, and $P\left((I, D_{init}) | D'\right)$ and $P(D')$ are the likelihood and prior probabilities, respectively. Taking the logarithm of the product of the likelihood and prior, the Bayesian formulation corresponds to minimization of the energy function:

$$D = \arg\min_{D'} E_d + \lambda E_s \quad (6)$$

where $E_d$ is the data energy expressed by the likelihood $P\left((I, D_{init}) | D'\right)$, $E_s$ is a smoothness term derived from the prior $P(D')$, and $\lambda$ is a weight. To incorporate the marginal and conditional NSS distributions that we have measured and modeled, together with the spatial smoothness constraints, the Bayesian inference formulation can be re-written as

$$\begin{aligned} D &= \arg\max_{D'} P\left(\tilde{D}' \middle| (I, D_{init}), \tilde{I}\right) \\ &= \arg\max_{D'} P\left((I, D_{init}) \middle| \tilde{D}', \tilde{I}\right) P\left(\tilde{I} \middle| \tilde{D}'\right) P(\tilde{D}') \end{aligned} \quad (7)$$

where $\tilde{I}$ and $\tilde{D}'$ are the magnitudes of the Gabor filtered responses of $I$ and $D'$, respectively. Taking the logarithm of Eq. (7) yields

$$D = \arg\min_{D'} \left[E_{data} + \lambda_n(E_{NSS_c} + E_{NSS_m}) + \lambda_s E_{smooth}\right] \quad (8)$$

where $E_{data}$ is the data energy derived from $P\left((I, D_{init}) \middle| \tilde{D}', \tilde{I}\right)$, $E_{NSS_c}$ and $E_{NSS_m}$ are energy terms related to the conditional and marginal NSS distributions, respectively, $E_{smooth}$ is the smoothness term related to depth changes, and $\lambda_n$ and $\lambda_s$ are constant weights.

All color images are first converted into the perceptually relevant CIELAB color space, then decomposed by the

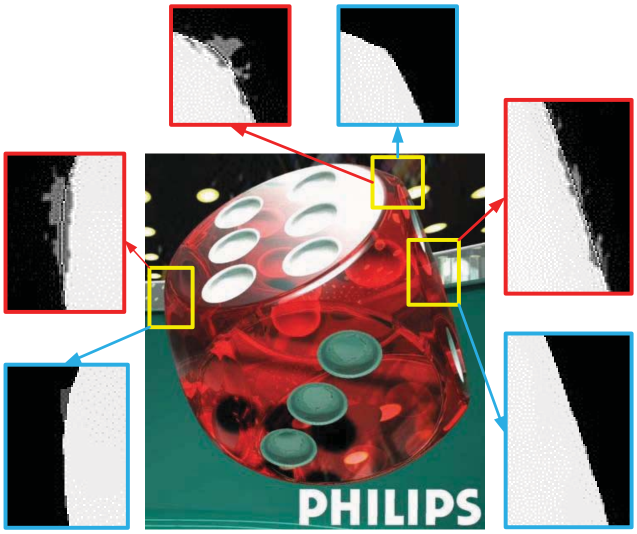


Fig. 5. Magnified view of depth estimation results obtained on the sequence "The Dice-1", the *red* rectangular zoomed areas are the depth estimated using the NSS-based method with just luminance, while the *blue* rectangular zoomed areas are the depths estimated using NSS models on the full color space.

afore mentioned Gabor filter bank. The Gabor magnitude response of each channel (depth, and $L^*$, $a^*$, $b^*$ in CIELAB color space) is then found. As both conditional and marginal terms are summed over the response of every filter in the Gabor filter bank, Eq. (8) can be re-written as

$$D = \arg\min_{D'} \sum_{i,j} \Big[ E_{data} + \lambda_s E_{smooth} + \sum_{f\in GB} \big( \lambda_m E_{NSS_{m,f}} + \sum_{k\in\{L^*,a^*,b^*\}} \lambda_k E_{NSS_{c,f,k}} \big) \Big] \tag{9}$$

where $GB$ is the set of filters in the Gabor filter bank.

Fig. 5 illustrates the effect of extending the natural scene statistics from luminance space to full color space, from which we can see that optimization using a chromatic natural scene statistics model yields more accurate depth values, especially around object edges.

## C. Data and Smoothness Terms

The data and smoothness terms here are similar to those used in classic stereo matching formulations. In a global optimization framework, the initial depth map serves as a reference and dramatically reduces the time cost. The data term is defined as

$$E_{data} = |D(i,j) - D_{init}(i,j)| \tag{10}$$

The smoothness term constrains the depth variations between adjacent pixels and thus forces the final solution to be smooth:

$$E_{smooth} = \sum_{(u,v)\in N(i,j)} \min\left(T_s, \left|D'(i,j) - D'(u,v)\right|\right) \tag{11}$$

where $T_s$ is the truncation threshold and $N(i,j)$ is a neighbourhood surrounding pixel $(i,j)$. Currently a 4-connected region is used.

## D. Distribution Terms

The generalized log-normal distribution effectively captures the shapes of the marginal empirical distributions of the magnitude Gabor responses to all four types of data. It is given by

$$p_g(x) = \begin{cases} \frac{\beta_g}{2x\alpha\Gamma\left(\frac{1}{\beta_g}\right)} \exp\left[-\left(\frac{|\ln(x)-\mu_g|}{\alpha_g}\right)^{\beta_g}\right], & x \geq 0 \\ 0, \quad x < 0 \end{cases} \tag{12}$$

where $\Gamma(\cdot)$ is the gamma function, and $\mu_g$, $\alpha_g$ and $\beta_g$ are location, scale and shape parameters, respectively.

Incorporating Eq. (12) into Eq. (9) yields

$$E_{NSS_{c,f,k}} = \ln\left(\tilde{I}_{f,k}(i,j)\right) + \ln\left(\frac{2\alpha_{f,k}\Gamma\left(\frac{1}{\beta_{f,k}}\right)}{\beta_{f,k}}\right) + \left(\frac{\left|\ln\left(\tilde{I}_{f,k}(i,j)\right) - \mu_{f,k}\right|}{\alpha_{f,k}}\right)^{\beta_{f,k}} \tag{13}$$

$$E_{NSS_{m,f}} = \ln\left(\tilde{D}'_f(i,j)\right) + \ln\left(\frac{2\alpha_{f,d}\Gamma\left(\frac{1}{\beta_{f,d}}\right)}{\beta_{f,d}}\right) + \left(\frac{\left|\ln\left(\tilde{D}'_f(i,j)\right) - \mu_{f,d}\right|}{\alpha_{f,d}}\right)^{\beta_{f,d}} \tag{14}$$

where $\mu_{f,k}$, $\alpha_{f,k}$, and $\beta_{f,k}$ are the location, scale, and shape parameters, respectively, of the best-fit generalized log-normal distributions of filtered luminance and chrominance conditioned on filtered depth, $\mu_{f,d}$, $\alpha_{f,d}$, and $\beta_{f,d}$ are the location, scale, and shape parameters of the best-fit generalized log-normal distribution of filtered depth, respectively, and $\lambda_k$ and $\lambda_m$ are their corresponding constant weights. What differentiates "conditional" from "marginal" is the way that we estimate the parameters of the generalized log-normal distribution. For the marginal distribution of depth, the parameters are taken directly from the natural scenes statistic results. For the conditional distribution of CIELAB, the parameters are further linearly modeled using the depth Gabor magnitude responses:

$$\mu_{f,k} = m_{\mu,f,k}\tilde{D}'_f(i,j) + b_{\mu,f,k} \tag{15}$$

$$\alpha_{f,k} = m_{\alpha,f,k}\tilde{D}'_f(i,j) + b_{\alpha,f,k} \tag{16}$$

$$\beta_{f,k} = m_{\beta,f,k}\tilde{D}'_f(i,j) + b_{\beta,f,k} \tag{17}$$

where $m_{\mu,f,k}$, $m_{\alpha,f,k}$, and $m_{\beta,f,k}$ are slope parameters for $\mu_{f,k}$, $\alpha_{f,k}$, and $\beta_{f,k}$, respectively, and $b_{\mu,f,k}$, $b_{\alpha,f,k}$, and $b_{\beta,f,k}$ are the corresponding offset parameters.

## E. Optimization and Discussions

To solve the optimization of the proposed Bayesian inference algorithm, we apply simulated annealing on the derived energy function [30]. The optimization process is started with an arbitrary assignment of depth. In the optimization of the target energy function, both the Gabor filter and the smoothness term are crucial to the final depth map quality, but their roles play differently. the effect of Gabor filter reflects in the parameters of generalized log-normal distribution via Eq. 15 to Eq. 17, and thus determines the penalty on depth

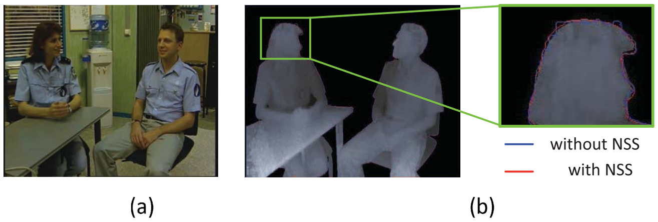


Fig. 6. Example of maintaining depth edges using NSS. The sequence used is "Interview", captured by Heinrich-Hertz-Institut. (a) Color frame; (b) Ground truth depth map, where red line shows the depth estimation result using NSS regularization while the blue line is without NSS.

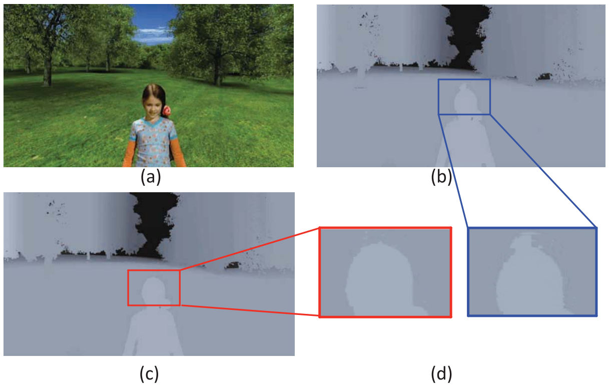


Fig. 7. Example of reduction of depth outliers using NSS. (a) Color image from the test sequence; (b) Depth map using the method of Li *et al.* [23] method without NSS regularization; (c) Depth map by the proposed method with NSS prior; (d) Magnified comparison.

changes in the target energy function. Large Gabor magnitudes attenuate the penalty of depth changes, while small Gabor magnitudes intensify the penalty. As a result, the main role of the Gabor filter in the proposed method is to link the color space and depth map so that the edges of depth changes can be found effectively. The Gabor filter is capable of detecting depth changes through its frequency and orientation representations, however, the depth edges it estimates from color space are not accurate enough for 3D rendering, which is very sensitive to object boundaries. Consequently, smoothness term is needed to guarantee the piecewise smooth within depth domain.

The incorporation of NSS confers two advantages. First, depth edges are strengthened along object boundaries as examplified by Fig. 6. If only Gabor filtering were applied, incorrect depth edges could be detected since there are color changes in the background Venetian blinds. The other advantage brought by NSS is a reduction of depth outliers remove, imposed by NSS regularization, as shown in Fig. 7.

## V. Experimental Results

In order to verify the effectiveness of the proposed scheme, we carried out both subjective and objective quality assessment to compare our method with several state-of-the-art 2D-to-3D video conversion methods (note that the comparison focuses on previous semi-automatic methods because in most cases, semi-automatic 2D-to-3D conversion has much better performance than its fully-automatic counterpart). The methods that are compared in the performance evaluation include: 1) bilateral filtering (BF) [31]; 2) improved depth propagation (IDP) [16]; 3) shifted bilateral filtering (SBF, bilateral filtering with temporal information) [18]; 4) bi-direction motion estimation and compensation (Li *et al*) [23]; and 5) The proposed NSS method. In order to illustrate the effect of each step for the proposed NSS-based 2D-to-3D conversion method, we also show the quality assessment result for natural scene statistics using only luminance (NSS L*); natural scene statistics using full color space with CIELAB (NSS L*a*b*); and natural scene statistics using L*, a*, b* with spatial smoothness (SS) regularization (NSS L*a*b* + SS) in Table II.

The test set consists of 10 different sequences, including various video clips ranging from indoor scenes to outdoor scenes, from computer graphics (CG) synthesized shots to real captured data. Sequences 1-8 were collected from the Philips WowVx© project website. Sequence 9 "Interview" was published by Heinrich-Hertz-Institut and sequence 10 "InnerGate" was synthesized by ourselves. The datasets are either synthesized by CG (Computer Graphics with ideal camera model) or captured by a camera that senses co-registered RGB and depth simultaneously, and thus can be regarded as ground truth data. These sequences contain challenging factors such as sharp edges, textureless regions, color ambiguity and objects having fast movement. The key frame interval was set in the range 20 to 30 frames for each sequence. Details and properties of the test sequences are listed in Table I. All of the test sequences can be downloaded from our website.[1] Thumbnails of some of the test sequences are shown in Fig. 8.

Comparisons of the converted depth maps on some of the test sequences are shown in Fig. 8. The first column is a frame from the sequence "Philips-3D-experience-1", the second column is a frame from the sequence "Dice-2", and the third column is a frame from the sequence "Interview". From this figure it may be seen that incorporating both natural scene statistics in full color space along with spatial smoothness results in the best depth estimates. Owing to limited space, not all of the estimated depth maps are shown here, but they are all available at our website.[2] Regarding evaluation of all the data sets from the various 2D-to-3D conversion methods, we performed both objective and subjective quality assessment as follows.

Computing the Mean Squared Error (MSE) between the degraded signal and the original ground truth is a straight forward way to measure the quality of a converted video. In the objective assessment part, we first used the MSE to measure the differences between the propagated depth maps and the ground truth. Table II shows the MSE comparison results. From the results we can see that if NSS is only applied on the luminance channel, the converted video quality is not always the best (see sequence 1, the previous SBF method outperforms the NSS using luminance). However, if we extend the NSS model into the full color space, the proposed conversion method delivers the best performance

[1] http://media.au.tsinghua.edu.cn/2Dto3D/testsequence.html

[2] http://media.au.tsinghua.edu.cn/2Dto3D/evaluation.html

TABLE I

SEQUENCES AND THEIR PROPERTIES USED IN OUR EXPERIMENTS. CG REPRESENTS COMPUTER GRAPHICS SYNTHESIZED VIDEO

| no. | sequence | generation method | video property |
|---|---|---|---|
| 1 | Inition-2d3d-Showreel-1 | real capture | outdoor scene, many textureless regions |
| 2 | Inition-2d3d-Showreel-2 | real capture | outdoor scene, fast movement objects |
| 3 | Philips-3D-experience-1 | CG + real capture | outdoor scene, textureless regions and sharp edges |
| 4 | Philips-3D-experience-2 | CG + real capture | outdoor scene, thin objects |
| 5 | Dice-1 | CG | indoor scene, color ambiguity and sharp edges |
| 6 | Dice-2 | CG | indoor scene, textureless regions and sharp edges |
| 7 | HeadRotate | real capture | indoor scene, occlusion and color ambiguity |
| 8 | Building | CG | indoor scene, color ambiguity and sharp-edges |
| 9 | Interview | real capture | indoor scene, repeatable textures |
| 10 | InnerGate | CG | outdoor scene, fast camera movement, zoom in/out |

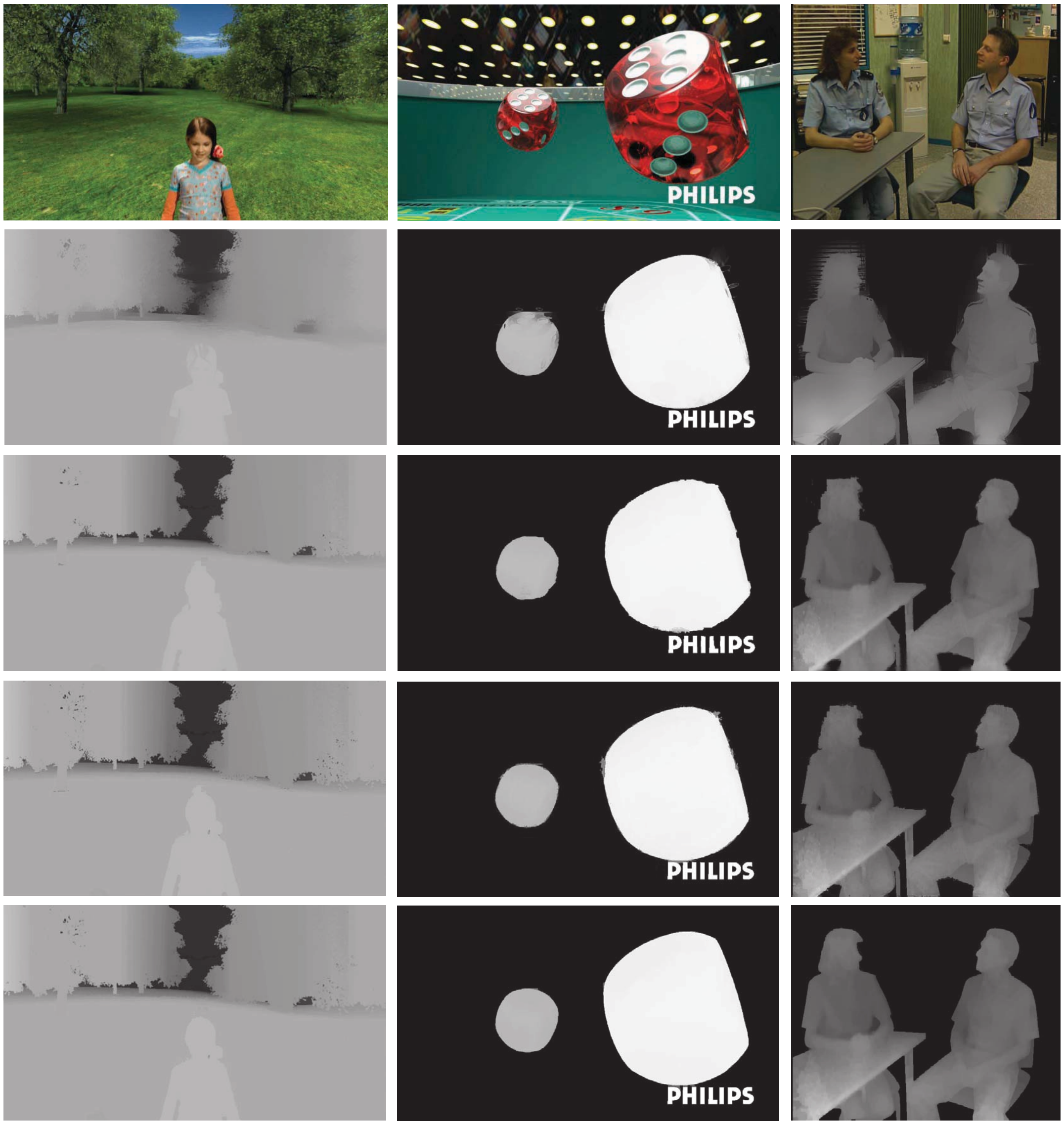


Fig. 8. Comparison of converted depth maps from different 2D-to-3D conversion methods. First row: color frame; second row: depth maps from the IDP method [16]; third row: depth estimation using the shifted bilateral filtering method [18]; fourth row: depth maps recovered using NSS with luminance [9]; last row: depth estimation using the proposed method.

on all the sequences. This agrees with the finding in [32] that color is an important ingredient in human stereopsis. Moreover, if the spatial smoothness term is also incorporated into the target function, the optimized depth values are improved even further because both the perceptual depth-color relationship and the spatial coherence of depth are considered. The last row of Table II shows the average MSE value for all the test sequences, from which we could see the NSS method gain much improvement in terms of objective evaluation score.

We also used the perceptually relevant Structure Similarity (SSIM) index to assess the 2D-to-3D converted

TABLE II

MEAN SQUARED ERROR (MSE) COMPARISON RESULTS

| No. | BF[31] | IDP method[16] | SBF[18] | Li *et al.*[23] | The Proposed NSS Method | | |
|---|---|---|---|---|---|---|---|
| | | | | | NSS with only L*[9] | NSS L*a*b* | NSS L*a*b* + SS |
| 1 | 42.59 | 40.91 | 47.46 | 16.89 | 17.07 | 15.54 | **14.13** |
| 2 | 7.76 | 7.55 | 8.51 | 5.51 | 5.25 | 4.93 | **4.87** |
| 3 | 87.13 | 94.83 | 40.04 | 41.98 | 37.08 | 30.34 | **26.87** |
| 4 | 607.15 | 548.94 | 245.50 | 190.77 | 177.20 | 165.99 | **120.34** |
| 5 | 131.40 | 124.75 | 249.98 | 86.97 | 58.75 | 57.15 | **55.10** |
| 6 | 71.18 | 70.01 | 191.54 | 69.25 | 41.72 | 41.24 | **40.55** |
| 7 | 85.70 | 79.78 | 40.58 | 19.27 | 18.57 | 17.94 | **17.09** |
| 8 | 387.67 | 360.47 | 227.29 | 105.81 | 89.31 | 86.57 | **84.01** |
| 9 | 112.32 | 98.93 | 68.23 | 45.03 | 40.57 | 34.17 | **31.76** |
| 10 | 497.47 | 529.96 | 400.77 | 156.41 | 123.65 | 116.81 | **114.52** |
| Avg. | 203.037 | 195.613 | 151.99 | 73.789 | 60.917 | 57.068 | **50.924** |

TABLE III

STRUCTURE SIMILARITY (SSIM) COMPARISON RESULTS. AVG REPRESENTS THE AVERAGE SSIM SORE

| No. | BF[31] | IDP method[16] | SBF[18] | Li *et al.*[23] | The Proposed NSS Method | | |
|---|---|---|---|---|---|---|---|
| | | | | | NSS with only L*[9] | NSS L*a*b* | NSS L*a*b* + SS |
| 1 | 0.967 | 0.971 | 0.974 | 0.979 | 0.978 | 0.981 | **0.982** |
| 2 | 0.984 | **0.985** | **0.985** | 0.981 | 0.982 | 0.983 | **0.985** |
| 3 | 0.961 | 0.971 | 0.977 | 0.976 | 0.976 | 0.978 | **0.980** |
| 4 | 0.912 | 0.928 | 0.933 | 0.935 | 0.931 | 0.937 | **0.947** |
| 5 | 0.978 | **0.988** | 0.978 | 0.985 | 0.984 | 0.987 | **0.988** |
| 6 | 0.983 | 0.990 | 0.981 | 0.987 | 0.986 | 0.990 | **0.991** |
| 7 | 0.973 | 0.976 | 0.981 | 0.987 | 0.987 | **0.988** | **0.988** |
| 8 | 0.840 | 0.875 | 0.902 | 0.922 | 0.921 | 0.922 | **0.928** |
| 9 | 0.951 | 0.963 | 0.976 | 0.979 | 0.978 | **0.984** | **0.984** |
| 10 | 0.891 | 0.900 | 0.913 | 0.930 | 0.935 | 0.936 | **0.937** |
| Avg | 0.944 | 0.955 | 0.960 | 0.966 | 0.967 | 0.969 | **0.971** |

TABLE IV

SUBJECTIVE TEST CONDITIONS

| | |
|---|---|
| 3D display | SONY LMD-4251TD |
| Display size | 42-inch (diagonal 107 cm) |
| 3D Display model | circular polarization |
| Peak luminance level | $70 \sim 250cd/m^2$ |
| Black luminance level | $< 0.7cd/m^2$ |
| Background room illumination | < 20 lux |
| Viewing distance | 3H (consistent with BT.1788) |
| Ratio of the luminance of the screen, when displaying only black level in a completely dark room, to that corresponding to peak white | 0.01 |
| Ratio of luminance of background behind picture monitor to peak luminance of picture | 0.15 |
| Chromaticity of background | $D_{65}$ |

results. SSIM takes structural information into account, and can better describe changes in content structure. Depth changes are crucial to the final viewing 3D experience. A high SSIM value indicates that two images or depth maps have more similarity of structure. We used the SSIM implementation for the LIVE website [33]. The calculated SSIM values are listed in Table III. From the table it may be observed that the proposed method also delivers the best performance in terms of SSIM, indicating that it is able to deliver depth structure with higher fidelity.

By far the most important test of any 2D-to-3D conversion is its perceptual efficacy. Therefore we also conducted a subjective test to further evaluate the perceptual efficacy of the various 2D-to-3D conversion methods. We recruited 22 subjects (with different age, occupation, and so on). All the participants were screened using the standard Snellen Eye Test Chart for visual acuity, and the Randot test for stereo depth perception. Only subjects having normal (corrected) vision participated in the study. We rendered the stereoscopic videos using a depth image based rendering algorithm, then displayed the videos on a 3D display. The 3D display used in the subjective test is a SONY LMD-4251TD 3D monitor, which is a 42-inch professional High-Definition display with circular polarization. The subjective test was performed according to the standard recommendations ITU-R BT.1438 [34] and ITU-R BT.1788 [35]. The viewing distance was about three times of the height of the display (160 cm), and the entire test for each subject was not more that 20 minutes. The subjective quality was recorded as scores considering the perceived sharpness, and overall depth of a set of stereoscopic image sequences [36]. In our experiments, video quality was rated on a continuous Mean Opinion Score (MOS) scale from $0 \sim 100$ (the viewer was asked to rate the quality of the video using any continuous number between $0 \sim 100$, the higher the number indicating higher perceived video quality). The viewing conditions are summarized in Table IV.

TABLE V

SUBJECTIVE MEAN OPINION SCORE (MOS) SCORES. AVG REPRESENTS THE AVERAGE MOS RESULT

| No. | BF [31] | SBF[23] | NSS with only L*[9] | The proposed NSS L*a*b* + SS |
|---|---|---|---|---|
| 1 | 70.60 | 73.00 | 76.60 | **77.05** |
| 2 | 78.60 | 76.40 | 78.85 | **79.05** |
| 3 | 79.85 | 79.40 | 78.80 | **85.05** |
| 4 | 80.10 | 82.65 | 86.15 | **86.20** |
| 5 | 79.10 | 81.90 | 81.65 | **83.40** |
| 6 | 66.90 | 75.35 | 75.40 | **76.60** |
| 7 | 71.55 | **76.40** | 76.35 | 75.90 |
| 8 | 72.60 | 74.55 | 72.40 | **78.30** |
| 9 | 75.70 | 72.95 | 79.15 | **79.60** |
| 10 | 72.00 | 75.05 | 78.35 | **80.50** |
| Avg | 74.70 | 76.77 | 78.37 | **80.17** |

The test results are shown in Table V. Our method delivered the best perceptual performance except on Sequence 7, where the results of the best algorithms were similar.

As for the computational cost, the proposed method requires 15 seconds on average to process one frame with 720 × 576 resolution, and requires 40 seconds on average to process one frame with High Definition.

## VI. CONCLUSIONS AND FUTURE WORK

We introduced a novel 2D-to-3D video conversion method inspired by psychophysical evidence of human visual processing of 3D scenes and by recent models of natural 3D scene statistics. The key contribution is that we designed a global depth optimization process that implicitly combines 2D color and 3D natural scene statistics with spatial depth coherence. Terms representative of these statistical and structural constraints in the designed target function serve as strong and effective priors on the optimization. The Bayesian inference framework makes it possible to force the 3D solution towards statistical naturalness and structural consistency. The algorithm derived from our model produces high-quality depth propagation over entire 2D video, leading to a better quality of experience when viewing 2D-to-3D converted content. The experimental results using both objective quality assessment indices and subjective experiments indicate that the proposed method delivers better performance than previous state-of-the-art 2D-to-3D conversion methods. Incorporating temporal "naturalness" and smoothness terms into the target energy function and seeking an effective energy minimization algorithm to accelerate the computation is a promising future direction.

In the future we intend to extend our work into the space-time modelling domain by using recent (temporal) video statistics models [37], [38]. Lastly, we also plan to explore methods of including content-dependence into the "naturalness" assessment process as in [39].

## REFERENCES

[1] L. Zhang, C. Vazquez, and S. Knorr, "3D-TV content creation: Automatic 2D-to-3D video conversion," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 372–383, Jun. 2011.

[2] X. Cao, A. C. Bovik, Y. Wang, and Q. Dai, "Converting 2D video to 3D: An efficient path to a 3D experience," *IEEE Multimedia*, vol. 18, no. 4, pp. 12–17, Apr. 2011.

[3] *Wikipedia*. [Online]. Available: http://en.wikipedia.org/wiki/2D_to_3D_conversion, accessed Nov. 18, 2014.

[4] R. W. Buccigrossi and E. P. Simoncelli, "Image compression via joint statistical characterization in the wavelet domain," *IEEE Trans. Image Process.*, vol. 8, no. 12, pp. 1688–1701, Dec. 1999.

[5] A. C. Bovik, "Automatic prediction of perceptual image and video quality," *Proc. IEEE*, vol. 101, no. 9, pp. 2008–2024, Sep. 2013.

[6] J. Portilla, V. Strela, M. J. Wainwright, and E. Simoncelli, "Image denoising using scale mixtures of Gaussians in the wavelet domain," *IEEE Trans. Image Process.*, vol. 12, no. 11, pp. 1338–1351, Nov. 2003.

[7] Y. Liu, L. K. Cormack, and A. C. Bovik, "Statistical modeling of 3D natural scenes with application to Bayesian stereopsis," *IEEE Trans. Image Process.*, vol. 20, no. 9, pp. 2515–2530, Sep. 2011.

[8] C. C. Su, L. K. Cormack, and A. C. Bovik, "Color and depth priors in natural images," *IEEE Trans. Image Process.*, vol. 22, no. 6, pp. 2259–2274, Jun. 2013.

[9] W. Huang, X. Cao, K. Lu, Q. Dai, and A. Bovik, "Towards naturalistic depth propagation," in *Proc. IEEE 11th IVMSP Workshop*, Jun. 2013, pp. 1–4.

[10] U. Rajashekar, Z. Wang, and E. P. Simoncelli, "Perceptual quality assessment of color images using adaptive signal representation," *Proc. SPIE*, vol. 7527, p. 75271L, Feb. 2010.

[11] M. Kim, S. Park, H. Kim, and I. Artem, "Automatic conversion of two-dimensional video into stereoscopic video," *Proc. SPIE*, vol. 6016, p. 601610, Nov. 2005.

[12] D. Kim, D. Min, and K. Sohn, "A stereoscopic video generation method using stereoscopic display characterization and motion analysis," *IEEE Trans. Broadcast.*, vol. 54, no. 2, pp. 188–197, Jun. 2008.

[13] A. Saxena, M. Sun, and A. Y. 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.

[14] 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.

[15] K. Karsch, C. Liu, and S. Kang, "Depth extraction from video using non-parametric sampling," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 775–788.

[16] C. Varekamp and B. Barenbrug, "Improved depth propagation for 2D to 3D video conversion using key-frames," in *Proc. 4th IETCVMP*, Nov. 2007, pp. 1–7.

[17] C. Wu, G. Er, X. Xie, T. Li, X. Cao, and Q. Dai, "A novel method for semi-automatic 2D to 3D video conversion," in *Proc. 3DTV Conf.*, May 2008, pp. 65–68.

[18] X. Cao, Z. Li, and Q. Dai, "Semi-automatic 2D-to-3D conversion using disparity propagation," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 491–499, Jun. 2011.

[19] O. Wang, M. Lang, M. Frei, A. Hornung, A. Smolic, and M. Gross, "StereoBrush: Interactive 2D to 3D conversion using discontinuous warps," in *Proc. 8th Eurograph. Symp. Sketch-Based Interf. Modeling*, 2011, pp. 47–54.

[20] J. S. Boreczky and L. A. Rowe, "Comparison of video shot boundary detection techniques," *J. Electron. Image*, vol. 5, no. 2, pp. 122–128, 1996.

[21] R. Zabih, J. Miller, and K. Mai, "A feature-based algorithm for detecting and classifying scene breaks," in *Proc. ACM 3rd Int. Conf. Multimedia*, 1995, pp. 189–200.

[22] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, vol. 1. Sep. 1999, pp. 377–384.

[23] Z. Li, X. Cao, and Q. Dai, "A novel method for 2D-to-3D video conversion using bi-directional motion estimation," in *Proc. IEEE ICASSP*, Mar. 2012, pp. 1429–1432.

[24] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, nos. 1–3, pp. 7–42, 2002.

[25] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *J. Opt. Soc. Amer. A*, vol. 2, no. 7, pp. 1160–1169, 1985.

[26] R. L. De Valois, D. G. Albrecht, and L. G. Thorell, "Spatial frequency selectivity of cells in macaque visual cortex," *Vis. Res.*, vol. 22, no. 5, pp. 545–559, 1982.

[27] J. P. Jones and L. A. Palmer, "An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex," *J. Neurophysiol.*, vol. 58, no. 6, pp. 1233–1258, 1987.

[28] M. Clark and A. C. Bovik, "Experiments in segmenting texton patterns using localized spatial filters," *Pattern Recognit.*, vol. 22, no. 6, pp. 707–717, 1989. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0031320389900071

[29] A. C. Bovik, M. Clark, and W. S. Geisler, "Multichannel texture analysis using localized spatial filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 1, pp. 55–73, Jan. 1990.

[30] S. T. Barnard, "A stochastic approach to stereo vision," in *Proc. 5th Nat. Conf. Artif. Intell.*, Aug. 1986, pp. 676–680.

[31] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proc. 6th Int Conf. Comput. Vis.*, Jan. 1998, pp. 839–846.

[32] J. R. Jordan, III, W. S. Geisler, and A. C. Bovik, "Color as a source of information in the stereo correspondence process," *Vis. Res.*, vol. 30, no. 12, pp. 1955–1970, 1990. [Online]. Available: http://www.sciencedirect.com/science/article/pii/004269899090015D

[33] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. (2003). *The SSIM Index for Image Quality Assessment*. [Online]. Available: http://www.cns.nyu.edu/~lcv/ssim

[34] *Subjective Assessment of Stereoscopic Television Pictures*, document ITU Rec., BT.1438, 2000.

[35] *Methodology for the Subjective Assessment of Video Quality in Multimedia Applications*, document ITU Rec., BT.1788, 2007.

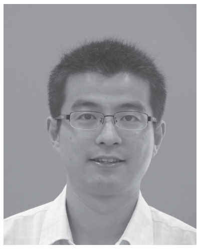
[36] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, document ITU Rec., BT.500, 1974–1997.

[37] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 4, pp. 684–694, Apr. 2013.

[38] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind prediction of natural video quality," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1352–1365, Mar. 2014.

[39] C. Li and A. C. Bovik, "Content-partitioned structural similarity index for image quality assessment," *Signal Process., Image Commun.*, vol. 25, no. 7, pp. 517–526, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0923596510000354

**Weicheng Huang** received the B.E. degree in automation from Tsinghua University, Beijing, China, in 2011, and the M.C.A. degree from the University of Chinese Academy of Sciences, Beijing, China, in 2014. His current research interests include 2D-to-3D conversion and video quality assessment.



**Xun Cao** (S'10–M'12) received the B.S. degree from Nanjing University, Nanjing, China, in 2006, and the Ph.D. degree from the Department of Automation, Tsinghua University, Beijing, China, in 2012. He is currently an Associate Professor with the School of Electronic Science and Engineering, Nanjing University. He visited Philips Research, Aachen, Germany, in 2008, and Microsoft Research Asia, Beijing, from 2009 to 2010. He was a Visiting Scholar with the University of Texas at Austin, Austin, TX, USA, from 2010 to 2011. His research interests include computational photography, image-based modeling and rendering, and 3DTV systems.

**Ke Lu** received the B.S. degree from the Department of Mathematics, Ningxia University, Yinchuan, China, in 1993, and the M.S. and Ph.D. degrees from the Department of Mathematics and the Department of Computer Science, Northwest University, Xi'an, China, in 1998 and 2003, respectively. He was a Post-Doctoral Fellow with the Institute of Automation, Chinese Academy of Sciences, Beijing, China, from 2003 to 2005. He is currently a Professor with the University of Chinese Academy of Sciences, Beijing. His research focuses on curve matching, 3D image reconstruction, and computer graphics.

**Qionghai Dai** (SM'05) received the Ph.D. degree in automation from Northeastern University, Shenyang, China, in 1996. He has been a faculty member since 1997 and a Professor since 2005 with the Department of Automation, Tsinghua University, Beijing, China. He has authored over 120 conference and journal papers, and holds 67 patents. His current research interests include the areas of computational photography, computational optical sensing, and compressed sensing imaging and vision. His work is motivated by challenging applications in the fields of computer vision, computer graphics, and robotics.

**Alan Conrad Bovik** (S'80–M'81–SM'89–F'96) holds the Cockrell Family Endowed Regents Chair in Engineering with the University of Texas at Austin, Austin, TX, USA, where he is currently the Director of the Laboratory for Image and Video Engineering. He is a faculty member with the Department of Electrical and Computer Engineering and the Institute for Neuroscience. His research interests include image and video processing, computational vision, and visual perception. He has authored over 700 technical articles in these areas and holds several U.S. patents. His publications have been cited more than 35 000 times in the literature. He is an H-index of over 70, and listed as a Highly-Cited Researcher by Thompson Reuters. His several books include the companion volumes entitled *The Essential Guides to Image and Video Processing* (Academic Press, 2009). He was a recipient of a number of major awards from the IEEE Signal Processing Society, including the Society Award in 2013, the Technical Achievement Award in 2005, the Best Paper Award in 2009, the SIGNAL PROCESSING MAGAZINE Best Paper Award in 2013, the Education Award in 2007, the Meritorious Service Award in 1998, and a co-author of the Young Author Best Paper Award in 2013. He was named as a recipient of the Honorary Member Award of the Society for Imaging Science and Technology in 2013, received the SPIE Technology Achievement Award in 2012, and was the IS&T/SPIE Imaging Scientist of the Year in 2011. He was also a recipient of the Hocott Award for Distinguished Engineering Research from the Cockrell School of Engineering, University of Texas at Austin in 2008 and the Distinguished Alumni Award from the University of Illinois at Urbana-Champaign, Champaign, IL, USA, in 2008.

Dr. Bovik is a fellow of the Optical Society of America and the Society of Photo Optical and Instrumentation Engineers. He co-founded and was the longest-serving Editor-in-Chief of the IEEE TRANSACTIONS ON IMAGE PROCESSING from 1996 to 2002, and created and served as the first General Chair of the IEEE International Conference on Image Processing in Austin in 1994, along with numerous other professional society activities, including the Board of Governors of the IEEE Signal Processing Society from 1996 to 1998, an Editorial Board Member of the PROCEEDINGS OF THE IEEE from 1998 to 2004, and a Series Editor of *Image, Video, and Multimedia Processing* (Morgan and Claypool Publishing Company, 2003-present). He was also the General Chair of the Texas Wireless Symposium in Austin in 2014. He is a Registered Professional Engineer in the State of Texas and a Frequent Consultant to legal, industrial, and academic institutions.