Using Wavelet Transformation and Edge Detection to Generate a Depth Map from a Single Image

Der-Feng Huang, Yu-Hsiang Chen and Ting-Wen Huang

Department of Electrical Engineering, LungHwa University of Science and Technology, Taoyuan City, Taiwan, ROC

ABSTRACT
A depth map is essential for 3D display. Traditional depth maps are derived from multi-images. Producing a depth map from a single image is a more challenging problem and much more difficult than that of from multi-images. We propose a novel approach that combines wavelet transformation with edge detection to generate depth information from a single image. A depth map is then generated by depth prediction. Peak signal-to-noise ratio (PSNR), the structural similarity image measure (SSIM), and computation time are used to evaluate the performance of the proposed algorithm. The experimental results reveal that the PSNR and SSIM of the proposed algorithm are larger than those produced in the traditional approach. The computation time of the proposed algorithm is less than that produced by the traditional approach. Furthermore, 1000 test images and the corresponding depth maps are produced in this study. The average of the three metrics was found to be superior to those of the traditional approach. The statistic tests were also significant for PSNR and TIME between the two approaches. We have also written and files four patent applications with respect to this work.

KEYWORDS
Depth map; wavelet transformation; edge detection

ARTICLE HISTORY
Received 26 January 2017
Accepted 12 March 2017

1. Introduction
The success of three-dimensional (3D) video games and the online 3D virtual world emphasizes the demand for viewing experience beyond the passive watching of two-dimensional (2D) television programs. In a traditional imaging system, a 3D scene is projected by a 2D image sensor and depth information is lost. Many 3D TVs which give a 3D appearance using binocular disparity have a 2D–3D conversion function. Various techniques have been introduced for 2D–3D conversion, and many of them are based on depth image-based rendering (DIBR) [1–3]. DIBR is a technique used to render at least two images (for the left and right eyes) from a 2D image and a depth map. In a typical procedure, a depth map is first generated, in one of a number of different ways, and the 3D images are then generated by DIBR using the depth map.

Valencia and Dagnino [4] proposed a method for deriving a depth map from a single image using wavelet analysis and edge estimation based on Lipschitz exponents. The relative depth can be estimated based on the values of wavelet coefficients in the high-frequency bands. Based on this, Valencia et al. divided images into 16 × 16-pixel macroblocks. A macro block wavelet transform which generated 256 wavelet coefficients was then performed. Relative depth was estimated by counting the number of
non-zero coefficients. In an early effort to recover depth using focus/defocus clues, a deconvolution method in the frequency domain was introduced by Pentland [5] in 1987 to determine the amount of image blur. A method for depth estimation based on focus/defocus cue where the entropy of high-frequency sub-band of wavelet decomposition is regarded as the measure of blur is shown in [6]. Mishiba et al. [7] introduced a depth estimation method using a wavelet-based matching cost optimization to estimate sharp object boundaries. There is also commercial hardware such as Kinect which consisting of an infrared projector and camera and a special microchip that generates a grid from which the location of a nearby object in three dimensions can be ascertained [8].

There are also some documented patents. For example, patent CN102427539B [9] proposed a method in which the input image is first wavelet transformed and the high-frequency components are separated into three different blocks, foreground, midground, and background. The depth map is obtained after defocus of the blocks. Patent CN103049906B [10] reveals an image depth extraction method comprised of the following steps: (1) treatment of the original image by Gaussian blur processing to give N pieces of blurred image. (2) Detection of the original image and the blurred image of N frames is done; N ≥ 2 fuzzy parameter estimates. (3) Based on the edge image, each pixel is calculated at each edge of the corresponding Gaussian filter parameters. (4) Each pixel at the edge of N fuzzy parameters is analyzed to estimate its worth using a statistical method. (5) Based on the parameter estimates at the edges of each pixel, the depth of each pixel in the image to the edges of the sparse depth map is calculated. (6) An interpolation process is used to obtain a dense depth map. The method described in this invention, relative to prior art methods, allows high accuracy to be obtained from fuzzy parameters and depth values can be accurately calculated.

Edge detection information is also a clue for the depth map. To create 2D-plus-depth format content from images of limited depth of field (DOF), the focused objects in the foreground are sharp-edged and defocused objects in the background are blurred [11]. That is, the focused area contains high-frequency components and blurred area contains low-frequency components. The degree of blurring can be directly related to the spatial frequency described by wavelet transformation. It suggests there are more details and less blurring in the high-frequency bands with more energy, where the objects are closer to the camera. Accordingly, the relative depth of each pixel can be estimated based on the values of wavelet coefficients. The high-frequency bands can then be derived using wavelet analysis. After two stages of smoothing and scale manipulation, the depth map data with fewer errors can be used for 3D display [12].

The high-frequency components represent focused objects in images with limited DOF and their high-frequency wavelet sub-bands contain high energy. For images of limited DOF, objects in the image may not all be in focus [5]. Usually, the objects in the background are blurred and texture is smooth, while the main foreground focused objects have sharp edges and detailed texture. In other words, the high frequencies are retained in the foreground and greatly attenuated in the background. This suggests that the local spatial frequency is directly related to the degree of blurring, and thus to their relative distance from the camera. The high frequencies are described by the coefficients of the wavelet transform of the image. If the energy of the high-frequency wavelet bands is significant, this suggests the region has more detail and less blurring.

The authors have written and filed four patent applications: two for inventions patents and two for utility models. The two invention patents are ‘A method and apparatus for combining wavelet transformation with corner point detection to generate a depth map from a single image’ (application No. 105121680). And, ‘A method and apparatus for combining wavelet transformation and edge detection to generate a depth map from a single image’ (application No. 105121679). The two utility model patents are ‘An apparatus for combining wavelet transformation with corner points to generate a depth map from a single image’ (application No. 105210335). And, ‘An apparatus for combining wavelet transformation with edge detection to generate a depth map from a single image’ (application No. 105210334).

This paper is organized as follows: after this introduction, the second section describes the problem, the third section presents the proposed algorithm, the fourth section is the experimental section, which includes statistical analysis results. The conclusion is reached in the last section.

2. The Problem

To produce a depth map from a single image is a challenging problem and much more difficult than doing the same from multi-images. The essence of the problem is to make a depth map from data extracted from a given 2D image with limited DOF. A depth map is a grayscale image in which distance information for each pixel in the image which is represented by shades of gray. For example, the objective of the proposed algorithm would be the conversion of the image shown in Figure 1(a) to that shown in Figure 1(b). In this image, the insect is in front and the other subjects are in the background. The depth information is shown in the gradient layer on the right side of Figure 1(b). The color of the object from white to
black corresponds to the distant of the object from near to far. Without loss of generality, we use uppercase letters to denote the image and images with \( M \times N \) pixels are considered in this paper.

Three metrics were used to evaluate the performance of the proposed algorithm. The first was the peak signal-to-noise ratio (PSNR) [13], PSNR is expressed as a ratio between the maximum possible power of a signal and the power of corrupting noise. The PSNR is usually expressed in terms of logarithmic decibel scale as follows:

\[
\text{PSNR} = 10 \log_{10} \frac{255^2}{\text{MSE}},
\]

where \( \text{MSE} = \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} [X(m, n) - \hat{X}(m, n)]^2 \), \( X \) and \( \hat{X} \) represent the original image and the depth map, respectively.

The second metric used was the structural similarity image measure (SSIM) [14] (see Figure 2). The system diagram of the SSIM is shown in Figure 3. Suppose \( X \) and \( Y \) are the original image and the corresponding depth map, respectively. We define \( \mu_X = \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} X(m, n) \), \( \sigma_X = \frac{1}{MN-1} \sum_{m=1}^{M} \sum_{n=1}^{N} [X(m, n) - \mu_X]^2 \), and the covariance \( \sigma_{XY} = \frac{1}{MN-1} \sum_{m=1}^{M} \sum_{n=1}^{N} [X(m, n) - \mu_X][Y(m, n) - \mu_Y] \).

For luminance comparison

\[
l(X, Y) = \frac{2\mu_{XY} + C_1}{\mu_X^2 + \mu_Y^2 + C_1},
\]

where \( C_1 = (K_1 L)^2 \) with \( L \) is the dynamic range of the pixel values (255 for 8-bit grayscale images) and \( K_1 \ll 1 \) is a small constant to avoid the denominator approaches 0 when \( \mu_X^2 + \mu_Y^2 \) is very close to 0. The contrast comparison function takes a similar form

\[
c(X, Y) = \frac{2\sigma_{XY} + C_2}{\sigma_X^2 + \sigma_Y^2 + C_2},
\]

where \( C_2 = (K_2 L)^2 \) with \( K_2 << 1 \) is a small constant to avoid the denominator approaches 0 when \( \mu_X^2 + \mu_Y^2 \) is very close to 0. The structure comparison function is as follows:

\[
s(X, Y) = \frac{2\sigma_{XY} + C_3}{\sigma_X \sigma_Y + C_3},
\]

where \( C_3 = \frac{C}{2} \). In general, the SSIM is defined as follows:

\[
\text{SSIM}(X, Y) = l(X, Y)c(X, Y)s(X, Y),
\]

Figure 1. (a) Original image, (b) depth map.

Figure 2. Flow diagram for structural similarity image measure (SSIM).
The third metric used was computation time. It is known that high values of PSNR and SSIM indicate better depth map quality. For the same test, hardware and software computation should be as fast as possible to ensure efficiency.

3. The Proposed Approach

The proposed algorithm has the following steps (see Figure 3):

1. 2D wavelet transformation of the input image with a specific threshold to produce a binary image.
2. Canny edge detection with a specific threshold to produce a binary image.
3. Perform logic AND on the two images to produce a defocused image.
4. Defocused image depth prediction.
5. Production of the depth map.

Figure 4 shows the wavelet transformation procedure. The converted grayscale image first passes a high/low-pass filter and then decimates along the rows, the result image then passes high/low-pass filter which decimates along the columns. We have the resulting HH, HL, LH, LL sub-images after the wavelet transform procedure. The HH sub-image consists the high-frequency information of the original image and hence its high energy. A threshold is needed to produce a binary image which represents the contour of the main object. Note that the size of the

\[
\text{SSIM}(X, Y) = \frac{(2\mu_X\mu_Y + C_1)(2\sigma_{XY} + C_2)}{\mu_X^2 + \mu_Y^2 + C_1}(2\sigma_{XY} + C_2)(\sigma_X^2 + \sigma_Y^2 + C_3)
\]

where \(\alpha, \beta, \gamma > 0\). Without loss of generality, \(K_1 = 0.01\), \(K_2 = 0.03\), \(\alpha = \beta = \gamma = 1\), and \(C_1 = \frac{1}{2}\) are used in this study.

We have the SSIM for comparison:

\[
\text{SSIM}(X, Y) = \frac{(2\mu_X\mu_Y + C_1)(2\sigma_{XY} + C_2)}{(\mu_X^2 + \mu_Y^2 + C_1)(\sigma_X^2 + \sigma_Y^2 + C_2)}
\]

Figure 3. The proposed algorithm.

Figure 4. Wavelet transformation.
The sub-image should be converted to the original size by four times. First, we find the values $h[x], (x = 0, 1, 2, \ldots, 255)$ in a plot of the grayscale image. The smoothed plot $h_0(x)$ can be found by

$$h_0(x) = \sum_{t=0}^{255} h[t] \delta(x - t) \otimes g(t) \quad (7)$$

where $\delta(t)$ is the continuous-time impulse response function and $g(x_0) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-x_0)^2}{2\sigma^2}}$ is the Gaussian kernel.

The maximum value can be found by $\frac{dh_0(x)}{dx} = 0$ and $\frac{d^2h_0(x)}{dx^2} < 0$. A typical plot is shown in Figure 5 and the threshold can be found as shown in Figure 6. After some experimentation, the threshold was chosen as follows:

$$Th = \frac{a + 4b + c}{6}, \quad (8)$$

where $a$, $b$, and $c$ are the top three values of the smoothed plot. Figure 7 is a typical wavelet transformed image computed with a specific threshold.

---

**Figure 5.** (a) A typical original grayscale plot and (b) the corresponding smoothed plot.

**Figure 6.** Computation of the wavelet transform threshold.

**Figure 7.** Original image (a) and a typical wavelet-transformed image computed with a specific threshold (b).
The Canny edge detector [15] with a threshold median value of 0.33 was used and Figure 8 shows the result of Canny edge detection.

The construct of the defocused image is obtained by

\[ f(m,n) = \begin{cases} 
  g(m,n), & \text{if } c(m,n) \neq 0 \\
  0, & \text{otherwise} 
\end{cases}, \quad (9) \]

where \( f(m,n) \) is the defocused image, \( g(m,n) \) is the wavelet transformed image and \( c(m,n) \) is the Canny edge detected image. The defocused image of the example is shown in Figure 9.

The depth interpolation can be produced from the defocused image by the approach used by Levin et al. [16]. Further, to avoid the stair effect in the image, suppose \( I \) and \( \hat{I} \) is the original image and defocused image, respectively. We propose the coefficient \( \lambda \) from solving the following equation [16]

\[ E(I) = I^T Li + \lambda(I - \hat{I}) D(I - \hat{I}), \quad (10) \]

where \( D \) is a diagonal matrix with entry \( D_{ii} = 1 \) if the \( ii \)-th entry of the defocus image is not 0, \( L \) is a square matrix with entry

\[ \sum_{k(i,j) \in w_k} \left( \delta_{ij} - \frac{1}{|w_k|} \left( 1 + (I_i - \mu_k)(\Sigma_k + \frac{1}{|w_k|} U_3)^{-1}(I_j - \mu_k) \right) \right), \quad (11) \]

where \( \delta_{ij} \) is the Kronecker delta function, \( U_3 \) is a 3 × 3 identity matrix, \( \mu_k \) and \( \Sigma_k \) are the mean and variance of the color matrix \( w_k \), respectively, \( I_i \) and \( I_j \) are the \( i \)-th and \( j \)-th entry of the original image \( I \), respectively, \( \varepsilon \) is an normalized parameter, and \( |w_k| \) is the size of window \( w_k \).

The depth value \( d \) can be solved by the equation

\[ (L + \lambda D)d = \lambda \hat{D}d. \quad (12) \]

The resulting depth map is shown in Figure 10.

4. Experimental Results and Statistical Discussion

Photo images are adopted from Flick and Chun-Po Chen in FaceBook. We used a total 1000 images for the test database. We use the computation time, PSNR, and SSIM to evaluate the performance of the proposed approach and of that used by Zhuo and Sim [17] as well.

4.1. The Experimental Results

For illustration, we use test image1 and test image2. The corresponding depth maps are shown in Figures 11 and 12, respectively. It can be seen that the two depth maps are almost the same. The values of PSNR, SSIM, and computation time for test image 1 and test image 2 are shown in Tables 1 and 2, respectively. It can be seen that the values of the three metrics of the proposed algorithm are superior to those of the conventional method.

Figure 13 shows the proposed PSNR (solid line) and the conventional PSNR (dotted line) [17] and it can be seen that the proposed PSNR outperforms the traditional approach. Figure 14 is the proposed SSIM (solid line) and the conventional SSIM (dotted line) [17]. It can be seen that the SSIM by the proposed approach is almost the same as that of the traditional approach. Figure 15 illustrates the relative difference (proposed computation time - conventional computation time) between the proposed and the conventional approach [17].
Table 1. Comparison of PSNR, SSIM with computation time.

| Test image | PSNR | SSIM | Time (s) |
|------------|------|------|----------|
| Zhuo and Sim [17] | 4.4  | 0.0141 | 17.6 |
| Proposed algorithm | 8.6  | 0.0308 | 21.5 |

Table 2. The comparison of PSNR, SSIM, and computation time.

| Test image | PSNR | SSIM | Time (s) |
|------------|------|------|----------|
| Zhuo and Sim [17] | 6.6  | 0.0273 | 21.2 |
| Proposed algorithm | 13.7 | 0.0555 | 20.7 |

Table 3. The average PSNR/SSIM/TIME.

|               | PSNR    | SSIM    | TIME (s) |
|---------------|---------|---------|----------|
| Zhuo and Sim [17] | 27.8282 | 0.4911 | 20.5531 |
| Proposed algorithm | 34.7133 | 0.4927 | 18.0599 |

**Figure 10.** A depth map of the example.

**Figure 11.** (a) Test image 1, (b) depth map by Zhuo and Sim method [17], (c) Depth map by the proposed algorithm.

**Figure 12.** (a) Test image 2 (b) depth map by Zhuo and Sim method [17] (c) depth map by the proposed algorithm.
and average computation time are shown in Table 3. It can be seen that the proposed approach outperforms the conventional approach.

To further illustrate the comparison of performance of the two approaches. Let

\[ X_p = \{ X_{p1}, X_{p2}, \cdots, X_{p1000} \} \]

and average computation time are shown in Table 3. It can be seen that the proposed approach outperforms the conventional approach.

4.2. Statistical Analysis of the Experimental Result

For the 1000 test images, the average PSNR

\[ \text{PSNR} = \frac{1}{1000} \sum_{n=1}^{1000} 10 \log \frac{255^2}{\text{MSE}_n} \]  \hspace{1cm} (13)

and average SSIM

\[ \text{SSIM}(X, Y) = \frac{1}{1000} \sum_{n=1}^{1000} \text{SSIM}(X_n, Y_n) \]  \hspace{1cm} (14)

and

Figure 13. The proposed PSNR (solid line) and conventional PSNR (dotted line).

Figure 14. The proposed SSIM (solid line) and the conventional SSIM (dotted line).
challenging problem and much more difficult than that of from multi-images. In this study, an approach is presented that combines wavelet transformation with edge detection to generate depth information from a single image. We use PSNR, SSIM, and computation time to evaluate the performance of the proposed algorithm. The experimental results reveal that the three metrics of the proposed algorithm outperform those produced in a conventional approach. One thousand test images and their corresponding depth maps are produced in this study. The average of the three metrics was found to be superior to those of the traditional approach. The statistic tests were also significant for PSNR and TIME between the two approaches. We have also written and files four patent applications with respect to this work.

Acknowledgments

The authors thanks the image providers in the Flick and Mr. Chun-Po Chen for his photos in the FaceBook. The authors thank the reviewers for their efforts in providing comments for helping to correct the error and enhance the quality of this paper.

Disclosure Statement

No potential conflict of interest was reported by the authors.

5. Conclusion

Traditional depth maps are derived from multiple-images. Producing a depth map from a single image is a more challenging problem and much more difficult than that of from multi-images. In this study, an approach is presented that combines wavelet transformation with edge detection to generate depth information from a single image. We use PSNR, SSIM, and computation time to evaluate the performance of the proposed algorithm. The experimental results reveal that the three metrics of the proposed algorithm outperform those produced in a conventional approach. One thousand test images and their corresponding depth maps are produced in this study. The average of the three metrics was found to be superior to those of the traditional approach. The statistic tests were also significant for PSNR and TIME between the two approaches. We have also written and files four patent applications with respect to this work.

Table 4. The p-value of PSNR and TIME.

|       | $\sigma^2_p = \sigma^2_c$ | $\sigma^2_p \neq \sigma^2_c$ |
|-------|--------------------------|--------------------------|
| PSNR  | $9.3599 \times 10^{-203}$ | $3.9419 \times 10^{-196}$ |
| TIME  | $8.5791 \times 10^{-41}$ | $1.9105 \times 10^{-40}$ |

Figure 15. Relative difference between the proposed computation time and conventional computation time.

Table 4. The p-value of PSNR and TIME.

$X_p = \{X_{p1}, X_{p2}, \ldots, X_{p1000}\}$

be the results from the proposed and conventional approaches, respectively. It is known that $X_p$ and $X_c$ are independent and the conditions of the central limit theorem apply. Assume $\mu_p$ and $\mu_c$ are the means of $X_p$ and $X_c$, respectively. $\sigma_p^2$ and $\sigma_c^2$ are the variances of $X_p$ and $X_c$, respectively. We consider a testing hypothesis on $\mu_p - \mu_c$.

\[
H_0 : \mu_p - \mu_c = 0 \\
H_1 : \mu_p - \mu_c > 0 \tag{15}
\]

The $p$-value of PSNR and TIME are shown in Table 4. It shows that no matter the variances ($\sigma_p^2$ and $\sigma_c^2$ equal or unequal), the $p$-values are extremely small [18]. That is, the test statistics are significant and the null hypothesis $H_0$ must be rejected in favor of the alternative hypothesis $H_1$. This concludes that the PSNR and TIME of the proposed approach outperform the traditional approach.

ORCID

Der-Feng Huang  http://orcid.org/0000-0003-3955-6949
References

[1] Zhang L. Stereoscopic image generation based on depth images for 3DTV. IEEE Trans Broadcast. 2005;51(2):191–199.
[2] Ko J, Kim M, Kim C. 2D to 3D stereoscopic conversion: depth-map estimation in a 2D Single-Vles image. Proc SPIE. 2007;6696.
[3] Knorr S, Sikora T. An image-based rendering (IBR) approach for realistic stereo view synthesis of TV broadcast based on structure from motion. IEEE Int Conf Image Process. 2007;6:572–575.
[4] Valencia SA, Rodriguez-Dagnino RM. Synthesizing stereo 3D views from focus cues in monoscopic 2D images. Proc SPIE. 2003;5006:377–388.
[5] Pentland AP. A new sense for depth of field. IEEE Trans Pattern Anal Mach Intell. 1987;9:523–531.
[6] Sun Huadong, Zhao Zhijie, Jin Xuesong, et al. Depth from defocus and blur for single image, 2013 visual communications and image processing; Kuching (Malaysia). 2013. p. 1–5.
[7] Mishiba Kazu, Oyamada Yuji, Kondo Katsuya. Wavelet-based depth map estimation for light field cameras. IEEE 5th Global Conference on Consumer Electronics; Kyoto (Japan). 2016. p. 1–2.
[8] Zhiyong P, Di Chen H, Jiawen J, et al. A Method for converting two-dimensional video image to three-dimensional video image. CN Patent, 102, 427, 539B; 2014 Aug 27.
[9] Qian WH, Wei W, Yongbing Z, et al. An image depth extraction method. CN Patent, 103, 049, 906, 2015 Sept 30.
[10] Fehn C, Barre RDL, Pastoor S. Interactive 3-DTV – concepts and key technologies. Proc IEEE. Mar 2006;94(3):524–538.
[11] Lin YH, Chiang TW, Tsai T, et al. Fast 3D image depth map estimation using wavelet analysis. 2010 International Conference on Audio Language and Image Processing (ICALIP); Shanghai (China). 2010. p. 1160–1163.
[12] Malik AS, Choi T-S, Nisar H. Depth map and 3D imaging applications: algorithms and technologies; IGI-global, Hershey (PA); 2012.
[13] Wang Z, Bovik AC, Sheikh HR, et al. Image quality assessment: from error measurement to structural similarity. IEEE Trans Image Process. 2004;13(4):600–612.
[14] Canny J. A computational approach to edge detection. IEEE Trans Pattern Anal Mach Intell. PAMI-8(6); 2009:679–698.
[15] Levin A, Lischinski D, Weiss Y. A closed-form solution to natural image matting. IEEE Trans Pattern Anal Mach Intell. 2008;30(2):228–242.
[16] Zhuo S, Sim T. Defocus map estimation from a single image. Comput Anal Images Patterns. 2011;44(9):1852–1858.
[17] Montgomery DC, Runger GC. Applied statistics and probability for engineers. 2nd ed. London: Wiley; 1999.