# 2D to 3D Image Conversion Based on Classification of Background Depth Profiles

Guo-Shiang Lin[1], Han-Wen Liu[2], Wei-Chih Chen[2],
Wen-Nung Lie[2], and Sheng-Yen Huang[3]

[1] Dept. of Computer Science and Information Engineering, Da-Yeh University
168, University Rd., Dacun, Chang-Hua, 515, Taiwan, R.O.C.
khlin@mail.dyu.edu.tw
[2] Department of Electrical Engineering
National Chung Cheng University, Chia-Yi, 621, Taiwan, R.O.C.
ieewnl@ccu.edu.tw
[3] Reallusion Inc., New Taipei City, Taiwan
elvis@reallusion.com.tw

**Abstract.** In this paper, a 2D to 3D stereo image conversion scheme is proposed for 3D content creation. The difficulty in this problem lies on depth estimation/assignment from a mono image, which actually does not have sufficient information. To estimate the depth map, we adopt a strategy of performing foreground/background separation first, then classifying a background depth profile by neural network, estimating foreground depth from image cues, and finally combining them. To enhance stereoscopic perception for the synthesized images viewed on 3D display, depth refinement based on bilateral filter and HVS-based contrast modification between the foreground and background are adopted. Subjective experiments show that the stereo images generated by using the proposed scheme can provide good 3D perception.

**Keywords:** 2D to 3D image conversion, background depth profile, stereoscopic perception, depth cue estimation.

## 1 Introduction

3D video applications, such as 3D multimedia, 3DTV broadcasting, and 3D gaming, are getting more popular due to an incredible viewing experience compared with 2D video. Among them, the 3D digital frame is promising in near future's consumer electronics products. Nowadays in the market, its LCD panel has been manufactured in a size of 7 inches that can be viewed without glasses (i.e., naked eye). A traditional 2D color image, either raw or decoded data, can then be converted into a left-and- right or a multi-channel format so as to be viewed on 3D displays.

To have a capability of multi-view conversion, the depth information that originally does not exist in the 2D color images needs to be estimated. Then, the Depth Image

Y.-S. Ho (Ed.): PSIVT 2011, Part II, LNCS 7088, pp. 381–392, 2011.

Based Rendering (DIBR) technique can be used to render/synthesize stereo or multi-views. Currently, researchers have proposed several 2D to 3D conversion algorithms for static images [2,13-15,17] and dynamic videos [4-6,12], aiming to mitigate the insufficiency of 3D contents. Due to less depth cues (e.g., motion) that can be found compared to video, images' 2D to 3D conversion is much more challenging.

Recent researches about automatic depth estimation from 2D photographic images can be divided into two categories. The first one is depth from defocus/focus. S. K. Nayer and Y. Nakagawa [1] explored the relationship between the focus level and the object distance from the focused plane, called SFF (Shape from focus), to estimate the depths. This method demands multiple images captured with different focal lengths, which is beyond our discussion. V. P. Namboodiri and S. Chaudhuri [2] proposed a method to perceive the depth layers from a single defocused image, called DFD (Depth from defocus). They estimate the blurring degree of each pixel and use it for assigning the relative depth. The other category of 2D depth estimation is based on multiple depth cues. For example, in [13], Hough transform is used to detect the vanishing point as the geometric cue, by which an initial depth map can be constructed. The depth map is then refined based on the texture cues extracted from the image segmentation result. In [14], wavelets transform of luminance (Y) component is used to detect high frequency of the foreground objects. For pixels of high spatial frequency, the depth is assigned larger (i.e., nearer). Their method is however preferably applicable to close-up images. On the other hand, Liu et al. [15] excludes the computation of depth cues from texture, contrast, or motion vector, but adopts a semantic-based algorithm which analyzes each image into parts of sky, land, building, etc. and assigns depths according to the result of semantic classification. On the other hand, Philips company [3] analyzes the image content to fit a background depth model. Discrete Cosine Transforms (DCTs) of the horizontal and the vertical projection profiles are performed and then a classifier is used to determine a best fit model according to the transformed coefficients.

In view of the human visual system (HVS), 3D space extensity perceived by human beings is mainly contributed from a layered or structured background depth and the relative depth between the foreground and the background. Based on this concept, we propose in this paper a 2D to 3D image conversion algorithm that integrates the processes of foreground/background separation, relative depth estimation for foregrounds, classification of and combination with a structured background depth profile, and post processing. Note that our classification of background depth profile is based on features of local texture gradient and local edge direction, aiming to provide better classification than that based on DCT coefficients of projection profiles. Our scheme is more generic and then more suitable for the conversion of 2D images including indoor, outdoor, landscapes, portrayal, etc.

The remainder of this paper is organized as follows. Section II describes the proposed depth estimation algorithm. Section III elaborates details of post processing to enhance the perceived stereo quality. In Section IV, experiment results are given and finally Section V draws some conclusions.

## 2 Proposed Depth Estimation Algorithm

Our proposed image conversion algorithm is illustrated in Fig.1, which consists of two parts: depth estimation and post processing. First, a segmentation-based method is applied to extract the foregrounds. Then the foreground depth and the background depth profile are estimated separately; the former is based on multiple depth cue estimation, while the later is based on neural classification. To enhance the perceived depth on a stereoscopic display, the initial depth map is refined by using the color information (e.g., alignment of color and depth edges) and the relative contrast between the foreground and background regions are tuned based on HVS. Finally, the refined color and depth information are both used to synthesize the stereo image pair by depth image based rendering (DIBR) technique.

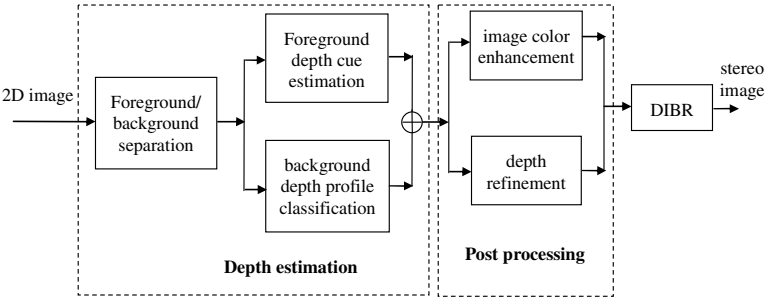



**Fig. 1.** The proposed 2D to 3D image conversion algorithm

### 2.1 Foreground/Background Separation

To extract the foreground regions, we adopt a strategy of performing region segmentation first and subsequently identifying regions that possibly belong to the foregrounds. There are several well-known region segmentation algorithms that have been proposed in literature. Among them, the mean-shift algorithm [10] is popularly used. Fig.2(b) demonstrates the segmentation result.

It is a challenging work to identifying foreground regions without some a priori knowledge. Based on an observation that foreground objects often occur at the central part of a frame (at least this assumption is valid for the digital frame application), we devise sampling boxes, as shown in Fig.2(b), to make statistics about the foreground (red) and background (green) color information. Since the pixel colors have been quantized by mean-shift algorithm, the results of color statistics will be limited. Denote the kinds of colors existing in the central and outer regions be $OC = \{oc_i | i = 1,...,M\}$ and $BC = \{bc_i | i = 1,...,N\}$, respectively. Our goal will be to delete from $BC$ the colors that possibly belong to foregrounds. Colors that retain in the revised $BC'$ will be used to extract the background regions.

Our method to classify the regions of a color $bc_i$ in $BC$ is to design filters based on a priori. A filter is used to sift out a color $bc_i$ from $BC$ if it also occurs in $OC$ and satisfies a certain criterion according to some region features (note that $bc_i$ may contain several disjointed regions in the frame). Region features are defined to include: position $(x, y)$ and size $(\Delta x, \Delta y)$ of the smallest enclosing rectangle, and compactness (the ratio between the region area size and $\Delta x \cdot \Delta y$ ). The criteria used in filters may be, e.g., the bottom $y$ of a background region should not be lower than a threshold (a lower region most probably belong to the foreground); regions of larger $\Delta x$ are possibly foregrounds. Fig.2(c) shows an example of foreground extraction (i.e., classifying regions of colors in $BC'$ as backgrounds and as foregrounds, otherwise). It can be seen that the result is satisfactory. Surely, the use of filters is not sufficient to sift out all false background colors in $BC$.

(a) (b) (c)

**Fig. 2.** (a) Original image, (b) result of mean-shift segmentation and the sampling boxes for the foregrounds (red) and the backgrounds (green), (c) identified foreground regions (shown with their mean-shift colors)

## 2.2 Foreground Depth Estimation

The cues for foreground depth estimation include texture gradient, sharpness, and face detection. Normally, a nearer object has stronger texture gradient and sharpness. However, these two cues are often indistinctive for the human face. We then adopt skin-color detection as an auxiliary tool to identify human faces as foregrounds. Note that since the depth image is usually smooth (i.e., of low spatial frequency), we calculate the depth cues based on blocks of 8×8 pixels to reduce the processing time.

(a) Texture gradient

The texture gradient of each pixel is calculated by using the Law's eight masks [7]:

$$z_i(x, y) = \left| \sum_{k=-1}^{1} \sum_{l=-1}^{1} w_i(k,l) I(x+k, y+l) \right| \tag{1}$$

where $I(x, y)$ is the intensity value at position $(x,y)$, and $w_i(k,l)$, $i$=1~8, denote the Law's masks. The texture gradient for each block is then defined as:

$$f^T(u,v) = \sum_{(x,y)\in block\_(u,v)} U\left(\sum_{i=1}^{8} z_i\,(x,y) - T_{t1}\right) \tag{2}$$

where $T_{t1}$ is a predefined threshold and $(u,v)$ is the block index.

(b) Sharpness

Empirically, edges of a near object have a sharper contrast than those of a far object. We define the variance and contrast of the graylevel in each block as the cues:

$$f^V(u,v) = \frac{1}{63}\sum_{(x,y)\in block\_(u,v)} \left(I(x,y) - \bar{I}_{u,v}\right)^2 \tag{3}$$

$$f^C(u,v) = \frac{I_{u,v}^{\max} - I_{u,v}^{\min}}{I_{u,v}^{\max} + I_{u,v}^{\min}} \tag{4}$$

where $\bar{I}_{u,v}$, $I_{u,v}^{\max}$, and $I_{u,v}^{\min}$ represent the average, maximum and minimum pixel values within the $(u,v)$-th block, respectively.

(c) Face cue

First, the input image is transformed from RGB to YCbCr color space. Pixels that satisfy conditions in both the RGB and YCbCr spaces are identified as the skin-color pixels [8]. Also, human's hair [9] (black is assumed) can be detected by using the algorithm proposed in [9]. The skin-color and hair-color pixels are united to form the human's information and assigned with a depth cue $f^P$=255; otherwise $f^P$=0.

(d) Depth cue fusion

Finally, the depth cues are fused to generate the depths for pixels located by the foreground mask obtained in Section 2.1 through Eq.(5):

$$f(u,v) = (w_1 \cdot f^V(u,v) + w_2 \cdot f^C(u,v) + w_3 \cdot f^T(u,v)) \cup f^P(u,v), \tag{5}$$

where $w_1 \sim w_3$ are predetermined weights ($w_1 + w_2 + w_3 = 1.0$) and "$\cup$" means pixel-wise maximum extraction. Since the depth cues are calculated in terms of blocks of 8×8 pixels, we apply a simple bilinear interpolation to rescale the foreground depth $f(u,v)$ to match the size of the input image.

### 2.3 Background Depth Profile Classification

We use a three-layer BPNN (Back-propagation Neural Network) to classify an image to one of the 5 types of background depth profiles. Figure 3 shows the 5 depth profiles defined in our system. The "1: up-bottom progressive" type is mostly often used in 2D-to-3D conversion. The "2: left-right progressive" and "3: right-left progressive"

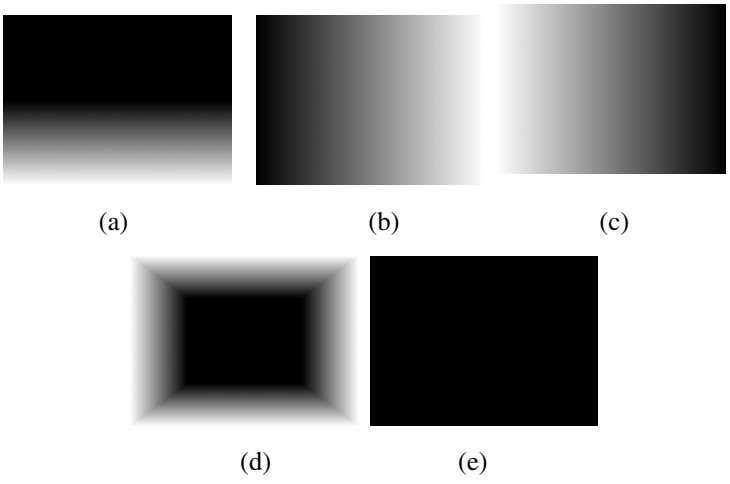



**Fig. 3.** Background depth profiles (a) up-bottom progressive, (b) left-right progressive, (c) right-left progressive, (d) indoor, and (e) close-up

show different increasing styles of depth. The "4: indoor" is most suitable for indoor scenario that constructs the strongest space extensity. On the other hand, "5: close-up" assumes the background depths all to zero, thus protruding the foreground objects substantially.

Features used in our neural network include:

1. local edge direction: The input image is divided into 3×3 regions, each is calculated the edge direction by using horizontal and vertical Sobel operator. All edge directions are quantized into 8 principle ones, each spaced by 45 degrees. The direction histograms of these 9 regions thus form the features of 72 dimensions.
2. local texture gradient: the average depth cues $f^T$ in these 9 individual regions, calculated based on Eq.(2), are also adopted as features of 9 dimensions.

Our algorithm is based on a similar observation in [16] that edge directions will have a dominant pattern which can be used to calculate the vanish point. Hence features based on local edge direction and texture gradient will be much more promising in practice. The three-layer neural network carried out to classify the background depth profile of each input image have 81 (72+9) input neurons, 50 hidden-layer neurons, and 5 output neurons. All input features are first normalized to be between 0.0 and 1.0. The output neuron with the highest score is selected as the background depth profile.

### 2.4 Combination of Foreground and Background Depths

The foreground and background depths obtained above are individually normalized to $(g^F_{\min}, g^F_{\max})$ and $(g^B_{\min}, g^B_{\max})$, respectively. These two sub-ranges can be fully or

partially overlapped (e.g., both are (0,255)), depending on user preferences. Finally, they are combined pixel-by-pixel by using a maximum operation. For non-foreground pixels, the final depth is the one calculated from the background depth profile; for foreground pixels, it is the maximum between foreground and background depths.

# 3 Color and Depth Post-processing

## 3.1 Depth Refinement

Since the depths are estimated at block-level first and then scaled up to the pixel-level, the depth edges may not be aligned with the color edges. This misalignment often causes quality degradation in the synthesized stereo images. We apply a bilateral filter [11] to refine the depth map. The bilateral filter is a weighted filter which evaluates the similarity of colors and distance between a current pixel and its neighboring ones, assigns the proper weights, and then calculates the weighted averages. It can not only smooth the depth map, but also make edges of foregrounds aligning to the color edges. Fig.4 (c) shows the refined depth map.

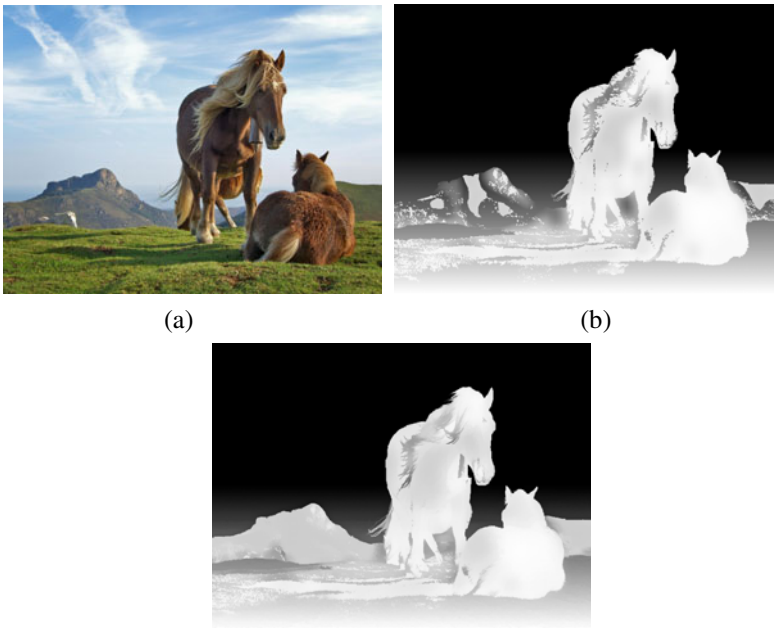



**Fig. 4.** Results of depth refinement (a) input image (b) initial depth estimation, (c) depth refined by bilateral filtering

### 3.2 Color Enhancement

It is known that the relative overlapping, foreground/background contrast, lighting, and shadows have influences on stereoscopic perception [12]. When looking at an image, people usually focus on foreground regions; the more the contrast between foreground and background regions, the more the stereoscopic perception. In this system, we apply two methods to modify colors of foreground/background pixels, according to the result of background depth classification, such that the stereoscopic effect is enhanced.

1. For background depth profiles #1-4, modify the RGB or hue-and-saturation (H/S) values of pixels in the *background t*o increase its contrast w.r.t. the foreground;
2. for background depth profiles #5 (i.e., close-up), modify the RGB or hue-and-saturation (H/S) values of pixels in *foreground* to increase its contrast w.r.t. the background.

Figure 5 demonstrates the original images and the color-enhanced images, respectively. It is seen that with proper color enhancement, the space extensity can be enhanced.

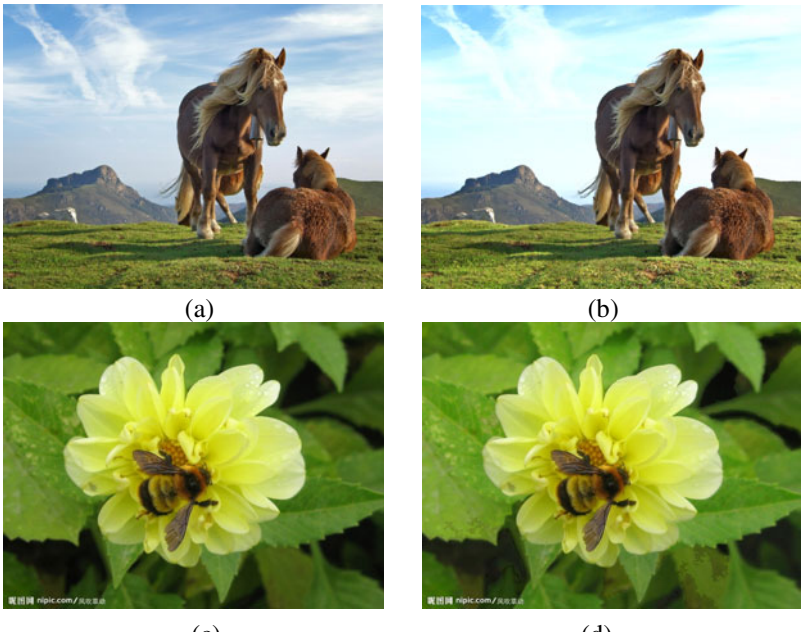



**Fig. 5.** (a)(c) Input images and (b)(d) color enhanced images

## 4 Experiment Results

The 3D display used in the experiments is an Acer 3D notebook (Model: 5738DG) equipped with polarizing glasses. Since depth estimation from a mono view is really challenging, we do not evaluate it by computing the objective quality metrics (e.g., PSNR) which require the existence of depth ground truths. Instead, subjective assessment based on Mean Opinion Score (MOS) of the synthesized stereo views is conducted. A total of 12 non-professional subjects are asked to score 1 to 5 (5 (excellent), 4 (good), 3 (fair), 2 (poor), and 1 (bad)) for each stereo pair generated by our proposed method (with $w_1$=0.4, $w_2$=0.2, and $w_3$=0.4). The test image size is all 640×480 pixels.

To evaluate background depth profile classification, 200 images (including landscape, portrait painting and indoor image, etc.) downloaded from the Internet are collected. Among them, the 5 types of background depth profiles are evenly distributed. A number of 75 images are used for training, 25 images for validation, and 100 images for testing. To determine the ground truths for neural network training, 5 subjects are asked to vote for the background depth profiles for each image. The dominant ones are selected as the truths. Table 1 shows the classification rates 91% and 83% for the training and test samples, respectively.

**Table 1.** Classification rates for background depth profiles

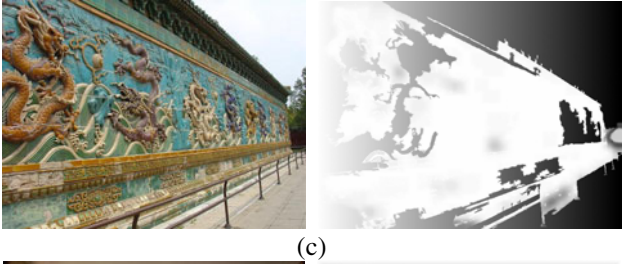
| Type of depth profile | Classification rate (training sample) | Classification rate (test sample) |
|---|---|---|
| 1 | 85 % | 80 % |
| 2 | 90 % | 100 % |
| 3 | 90 % | 85 % |
| 4 | 90 % | 70 % |
| 5 | 100 % | 80 % |
| Average | 91 % | 83 % |

Examples of some estimated depth maps are given in Fig. 6. Their background depth profiles are automatically classified as Type 1~5, respectively. The finally estimated depth maps are satisfactory.

Fig. 7 shows the MOS scores of the proposed algorithm for test images of different kinds of background depth profiles. Obviously, the close-up category yields the highest stereoscopic perception.

(a)

(b)



(c)

(d)

(e)

**Fig. 6.** (Left) input image (right) estimated depth map

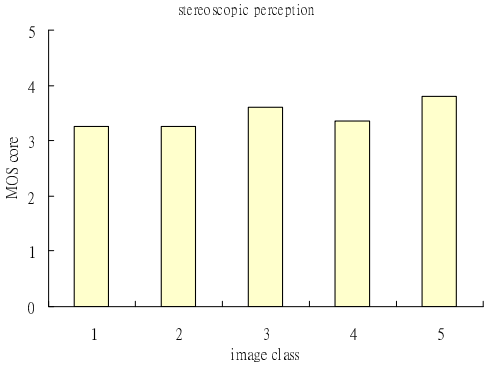



**Fig. 7.** MOS scores for test images of different kinds of background depth profiles

## 5 Conclusions

In this paper, we propose a 2D to 3D image conversion scheme. Our scheme is featured of: 1) segmentation-based foreground extraction, 2) foreground depth estimation based on multiple depth cue, 3) neural-network-based background depth profile classification, and 4) color enhancement for stereoscopic perception. Experiments show that our background depth classification has achieved a correct rate of 83% and the quality of synthesized stereo images viewed on the 3D display is good.

## References

1. Nayar, S.K., Nakagawa, Y.: Shape from Focus. IEEE Trans. on Pattern Analysis and Machine Intelligence 16(8), 824–831 (1994)
2. Namboodiri, V.P., Chaudhuri, S.: Recovery of Relative Depth from a Single Observation Using an Uncalibrated (Real-Aperture) Camera. In: Proc. of IEEE International Conference on Computer Vision and Pattern Recognition, Anchorage, AK, pp. 1–6 (2008)
3. Burazerovic, D., Vandewalle, P., Berretty, R.P.: Automatic Depth Profiling of 2D Cinema - and Photographic Images. In: Proc. of IEEE International Conference on Image Processing, Cairo, pp. 2365–2368 (2009)
4. Kim, M., Park, S., Kim, H., Artem, I.: Automatic conversion of two-dimensional video into stereoscopic video. In: Proc. of SPIE, vol. 6016, pp. 601610-1–601610-8 (2005)
5. Manbae, K., Sanghoon, P., Youngran, C.: Object-Based Stereoscopic Conversion of MPEG-4 Encoded Data. In: Proc. of the 5th Pacific-Rim Conference on Multimedia, pp. 491–498 (2004)
6. Kim, D., Min, D., Sohn, K.: A Stereoscopic Video Generation Method Using Stereoscopic Display Characterization and Motion Analysis. IEEE Trans. on Broadcasting 54(2), 188–197 (2008)

7. Suzuki, M.T., Yaginuma, Y., Yamada, T., Shimizu, Y.: A Shape Feature Extraction Method Based on 3D Convolution Masks. In: Proc. of Eighth IEEE International Symposium on Multimedia, pp. 837–844 (2006)
8. Peer, P., Kovace, J., Solina, F.: Human Skin Color Clustering for Face Detection. In: EU1ROCON 2003 International Conf. on Computer as a Tool, vol. 2, pp. 144–148 (2003)
9. Chen, Y.-J., Lin, Y.-C.: Simple Face-detection Algorithm Based on Minimum Facial Features. In: Proc. of IEEE International Conference on Industrial Electronics Society, pp. 455–460 (2007)
10. Sudhamani, M.V., Venugopal, C.R.: Segmentation of Color Images using Mean Shift Algorithm for Feature Extraction. In: Proc. of IEEE International Conference on Information Technology (2006)
11. Tomasi, C., Manduchi, R.: Bilateral Filtering for Gray and Color Images. In: Proc. of IEEE International Conference on Computer Vision, Bombay, pp. 839–846 (1998)
12. Lin, G.-S., Yeh, C.-Y., Chen, W.-C., Lie, W.-N.: A 2D to 3D conversion scheme based on depth cues analysis for MPEG videos. In: IEEE International Conference on Multimedia and Expo, pp. 1141–1145 (2010)
13. Han, K., Hong, K.: Geometric and texture cue based depth-map estimation for 2D to 3D image conversion. In: IEEE International Conference on Consumer Electronics, pp. 651–652 (2011)
14. Chiang, T.-W., Tsai, T., Lin, Y.-H., Hsiao, M.-J.: Fast 2D to 3D conversion based on wavelet analysis. In: IEEE International Conference on Systems Man and Cybernetics, pp. 3444–3448 (2010)
15. Liu, B., Gould, S., Koller, D.: Single image depth estimation from predicted semantic labels. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1253–1260 (2010)
16. Choi, J., Kim, W., Kong, H., Kim, C.: Real-time vanishing point detection using the local dominant orientation signature. In: 3DTV Conf., Turkey (May 2011)
17. Saxena, A., Sun, M., Ng, A.Y.: Learning 3-D Scene Structure from a Single Still Image. In: ICCV 2007 (2007)