Binocular Stereopsis of Traditional Chinese Paintings

Wei Ma*
Hae Kwang Kim†
Department of Computer Science
Sejong University

Yizhou Wang‡
Wen Gao§
Nat'l Engineering Lab. for Video Technology
State Key Lab. of Machine Perception (MOE)
Peking University

Weon Geun Oh¶
ETRI

Abstract

This paper presents an interactive technique for generating stereoscopic images from traditional Chinese paintings. The technique exploits a traditional way of moving focus based spatial composition used by classical Chinese painters. According to the moving focus rules of depicting objects of a painting, each non-ground pixel has the same depth with its corresponding pixel on the ground. Therefore, the depth map computation of the painting is decomposed into two steps. Firstly, the depth map of the ground is computed based on the moving focus rules of expressing spatial depth variation, with line-shape cues given through a user interface. Secondly, in order to calculate the depth of pixels belonging to the non-ground objects, the user interface provides a tool to sketch the objects and their occupying regions in the ground part. After that, each pixel of the non-ground objects is automatically matched to a pixel in the occupying ground regions by linear interpolation and takes the depth of the ground pixel as its depth. Finally, an anaglyph is computed by the obtained depth map. Experimental results demonstrate that the method presented in this paper can generate convincing binocular stereo images with easy user interaction.

CR Categories: I.4.9 [Image Processing and Computer Vision]: Applications—;

Keywords: stereopsis, traditional Chinese paintings, exhibition of paintings, e-heritage

Links: DL PDF

1 Introduction

As a new and promising digital media, stereo images/videos have gained much attention [Nedovič et al. 2010a] [Jobson et al. 2010]. These data can be obtained by capturing originals using stereo cameras, in case that the originals are 3D entities but 2D images. The only way of making existing 2D images to be stereo is inferring their depth/disparity maps by pictorial cues. There exist a number of methods to solve this problem [Nedovič et al. 2010a] [Jobson et al. 2010]. However, most of these work targets at photos or videos of natural scenes. This paper mainly focuses on the stereolization of traditional Chinese paintings (TCP for short) of landscape. As stated in [Rowley 1947], TCPs of landscape are much different from photos/videos, the products of light and cameras, which make it hard to apply the state-of-the-art methods to them.

This paper proposes a method to convert TCPs to stereoscopic images by exploiting a traditional way of moving focus based spatial composition used in TCPs [Fong 2003]. A user interface is developed to assist the conversion. Given an image of a painting, first, we interactively indicate extremely limited line-shape cues about the ground described in the image. An automatic computation procedure is then performed to obtain the depth map of the ground by the rules of moving focus. Next, the depth of pixels belonging to non-ground parts is computed by the depth map of the ground, based on the principles of moving focus and with the help from the user interface. Finally, an anaglyph-form stereoscopic image is generated by the depth map for display. Experimental results show that the proposed method can provide a different and interesting show of ancient paintings. To the best of our knowledge, we are the first to develop interactive stereolization techniques for TCPs.

The remaining of the paper is organized as follows. Section 2 reviews the state-of-the-art literatures related to depth inference for stereolization of single-view 2D images/videos. Section 3 describes the proposed method of converting TCPs to stereo in detail. Section 4 presents experimental results. Section 5 draws a conclusion.

2 Related work

Methods of depth inference from single view images/videos can be roughly classified into three categories, motion assisted, statistical based and interactive ones. Motion assisted approaches mainly target videos and use motion information to estimate depth. Here, the motion can be of the cameras or scenes. In the former case, scenes are generally static and their depth is computed by stereo matching [Zhang et al. 2009]. In the latter case, cameras are fixed while scenes have dynamic objects in them. The depth of a pixel is set to be proportional to the extent of its motion computed through neighbor frames [Kim et al. 2008] [Guttmann et al. 2009].

Statistical methods generally learn from a training set the relations between depth and image features [Nedovič et al. 2010a], depth and image semantic structures [Torralba and Oliva 2002], depth and image global structures [Nedovič et al. 2010b], or depth and semantic labels [Jobson et al. 2010]. The training data is generally restricted to have similar structures and appearances with the queries. The depth information in the training data is obtained by manual labeling [Torralba and Oliva 2002] or laser scanners [Nedovič et al. 2010a].

The two categories of methods above are not applicable in the task of this paper, since it is hard to extract motion information from paintings, or find structurally/apparently similar paintings with depth information for training.

*e-mail: rubbymawei@gmail.com
†e-mail:hkkim@sejong.ac.kr
‡e-mail:yizhou.wang@pku.edu.cn
§e-mail:wgao@pku.edu.cn
¶e-mail:owg@etri.re.kr

Interactive methods provide direct cues of depth through user interfaces. Two straightforward interactive methods are 1) determining the depth of each image pixel totally by manual, as done by using Photoshop [PSMethod1 2008] and 2) first segmenting objects out and then determining their depth object by object subjectively, as done by [PS2Method2 2010] in Photoshop. However, these methods are time consuming to get reasonable results. Intelligent methods make use of the inherent properties in photos. For example, Gimpel3D [Gimpel3D 2011], a free editor for 2D to stereo conversion, utilizes the perspective model of photos to facilitate the interactive conversion and generates convincing results. Guttmann et al. [Guttmann et al. 2009] labels the depth of a small part of an image, and then automatically propagate the depth labels to the other pixels in the same image or to temporal neighbors in videos. The propagation relies on the projective geometry of cameras, the appearance similarities between the labeled and unlabeled pixels, and a smoothness prior on depth. The method developed in this paper is also intelligent. Differently, the paper targets TCPs rather than photos.

3 The proposed method

The proposed method generates stereoscopic images from TCPs of landscape by exploiting the properties of the moving focus based spatial composition used in them. Spatial composition is the technique of positioning objects (or object parts) in 2D canvases for 3D spatial perception of viewers. In photos, the spatial composition is determined by perspective projection through optical imaging systems. In TCPs, the spatial composition technique is moving focus based [Fong 2003]. In contrast to the fixed single viewpoint in perspective, as stated in [Fong 2003], classical Chinese painters position objects (or object parts) on the picture plane additively, in an expanded field of vision, owing to his/her *moving focus* dynamic in parallel. More characteristics of the moving focus based spatial composition will be introduced when being used in the depth computation.

Considering that TCPs of landscape are composed of ground and non-ground objects supported by the ground [Yang et al. 1997], we perform the depth computation for the two parts sequentially.

3.1 Ground depth computation

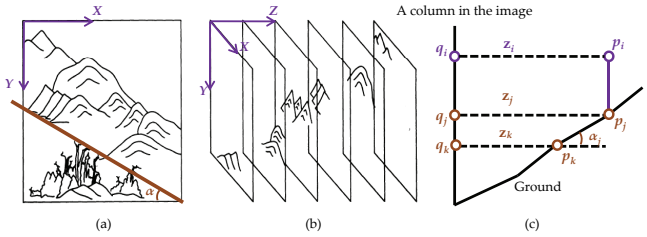


Figure 1: *(a) Part of a moving-focus painting with a sequence of objects. α is the inclination angle of the sequence. (b) Layers of the sequence. The background pictures of (a) and (b) are from [Fong 2003]. (c) Illustration of depth computation. See text for details.*

As described in [Fong 2003] about moving focus, in order to create the perception of varying depth of the ground on 2D canvas, painters align the bottom of a part of static objects, such as buildings and hills, in a sequence with a certain angle relative to the horizontal axis of the picture plane [Fong 2003], as illustrated in Figure 1(a)(the sequence of hills). The sequence also can be a single object positioned in a certain angle, such as the base of the tree and that of the rockery in Figure 2(a). We use a slanting line (the brown line in Figure 1(a)) to denote the sequence. The angle between the slanting line and the horizontal axis of the image plane is defined as the inclination angle (α in Figure 1(a)) of the sequence. The smaller the angle, the larger the depth variation of the sequence and the ground around. Long/high scrolls generally have groups of sequences with varying inclination angles.

A user interface is provided to interactively indicate the slanting lines. As shown in Figure 2(a), two lines are drawn along the direction of the tree base and that of the rockery base. The far background buildings can be treated as a single layer without expressing depth variation. Then the inclination angles are propagated to every pixel in the image. In this process, the pixels on the lines are treated as references, which have inclination angles of the lines as described above. For a pixel off the lines, its inclination angle is interpolated by its left and right reference pixels in the same row. For pixels having no references, such as those on the top part of Figure 2(a), they inherit the inclination angles of their nearest neighbors in the same column.

Figure 2: *(a) Line cues. (b) Computed ground depth.*

With the inclination angles reflecting depth variation, we proceed to compute the depth of the ground. The ground is globally distributed in the image, even though parts of it are occluded by the other objects. In this step, we assume that the occluding objects do not exist. Therefore, all the pixels in the image are treated as ground ones. The image coordinates system is defined as shown in Figure 1(a). The coordinates system of the 3D space has the same X and Y axes and the origin with the image coordinates system and a Z axis perpendicular to the image plane. Given a ground pixel $q_j = (x_j, y_j)$, its depth is defined as its distance from its corresponding point p_j on the 3D ground along the viewline through q_j (refer to Figure 1(c)). According to the definition of moving focus [Fong 2003], viewlines are perpendicular to the picture plane (as illustrated by the dash lines in Figure 1(c)). Therefore, p_j has the same X and Y components with q_j, and the depth of q_j is the Z component, z_j, of p_j.

The depth of the ground is computed row by row from bottom to top. Assuming q_j's bottom neighbor is $q_k = (x_k, y_k)$ ($x_k = x_j$ and $y_k = y_j + 1$) with already computed depth z_k (refer to Figure 1(c)), q_j's depth z_j is given by

$$z_j = z_k + ctan(\alpha_j), \tag{1}$$

where, α_j is the inclination angle of q_j. The computation is based on the fact that the depth variation is inversely proportional to the inclination angle as mentioned earlier. The depth of the pixels in the bottommost row equals zero since we assume that the 3D ground borders the painting, as illustrated in Figure 1(c). Using this method, the depth map of the ground is obtained as shown in Figure 2(b).

3.2 Depth map computation

In this subsection, we compute the depth of non-ground pixels by the ground depth map. According to the description about TCPs in

[Fong 2003], 1) objects (or object parts) are presented frontally and additively in parallel layers (as shown in Figure 1(b)). These layers are parallel to the image rather than perpendicular to the 3D ground plane for fully depicting the objects; 2) the artful overlapping and directional variation of strokes on a single layer create the illusion of continuous 3D space rather than discrete layers.

The algorithm of computing the depth of the non-ground pixels is designed based on the above facts. We interactively sketch a polygon encircling an object (or object part). This polygon is called layer polygon (e.g. the purple one encircling the women in Figure 3). Each layer polygon occupies a region on the ground in the image. According to the first fact, the occupying region of the layer is a line. Each pixel in the layer has the same depth with its vertical projection on the line, resulting in a planar depth map of the layer. On the other side, considering the second fact, we make the depth of the layer to be non-planar by indicating its occupation on the ground to be a region, in form of a polygon (e.g. the brown one at the foot of the women in Figure 3), rather than a line. For manmade objects, such as buildings, there exist layers not erect. In such case, we do the same for such layers as the the parallel ones. A user interface is provided for sketching the layer polygons and their occupying regions.

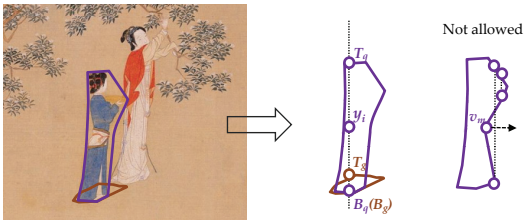


Figure 3: *Interpolation for finding corresponding pixels on the ground. See text for details.*

Given a pixel $q_i = (x_i, y_i)$ in a layer, its corresponding positions on the ground is found by interpolation along Y axis. The topmost point and the bottommost one of the layer polygon correspond to the topmost point and the bottommost one of the occupying polygon along the line $X = x_i$, as shown in Figure 3. Therefore, q_i's corresponding pixel on the ground $q_j = (x_j, y_j)$, in which $x_j = x_i$, and

$$y_j = B_g - (B_g - T_g)\frac{B_g - y_i}{B_q - T_q}. \quad (2)$$

Here, B_q, T_q, B_g, T_g are the bottommost and topmost points of the layer polygon and those of the occupying region, along $X = x_i$. The depth of q_i is set to be that of its corresponding pixel q_j. The layer depth computed in this way varies in vertical and horizontal directions, thanks to the occupying polygons. For a fine depth map, we recommend a pre-segmentation of each object (or object part) as done in the other interactive methods [PSMethod1 2008] [Gimpel3D 2011]. In such cases, only pixels really belonging to an object (or object parts), rather than lying in their layer polygon, are involved in the depth computation described above. However, for objects with concise contours, such as buildings, there is no need to do segmentation since the layer polygons can describe their contours precisely.

Note that 1) the occupying polygons should be wider than theirs corresponding layer polygons, so that any pixel in the layer polygons can find its corresponding pixel on the ground using the above method; 2) the layer polygons are restricted to having no sharp changes of the value $B_q - T_q$ and $B_g - T_g$ between neighbor columns as shown in Figure 3. Otherwise, there will be unsmooth depth transition between the columns according to Equation 2. Users can draw a polygon freely in the user interface. The interface will finally generate a layer/occupying polygon used for interpolation by moving parts of its vertices in parallel, e.g. v_m in Figure 3.

3.3 Anaglyph generation

With the depth map, the disparity d_i for each pixel is given by,

$$d_i = \frac{max_{z_i} - z_i}{max_{z_i}} \frac{w}{\delta}. \quad (3)$$

Here, z_i (quantized to be integers in [0,255]) is the depth of the pixel, max_{z_i} denotes the maximum value in the depth map, w is the width of the image, and $\delta = 20$ in the experiments. After that, we treat the original image as the left image and translate the pixels in it to the left with the extent of their disparities to generate a right image. The gaps appearing due to translating are filled with the color of the canvas. The red channel of the left image and the green and blue channels of the right image are merged to form an anaglyph.

4 Experimental results

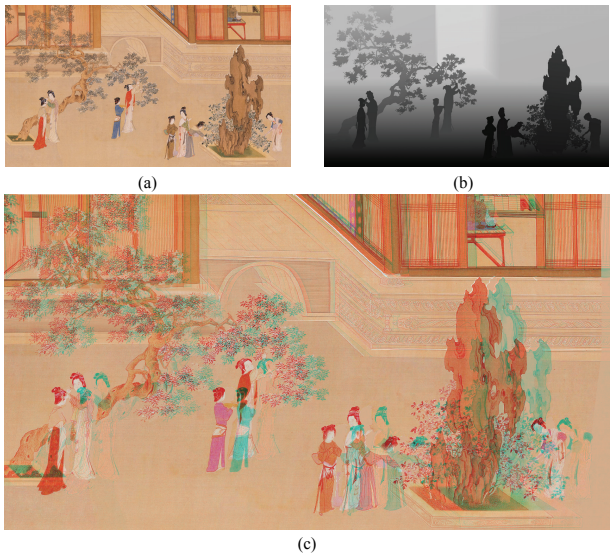
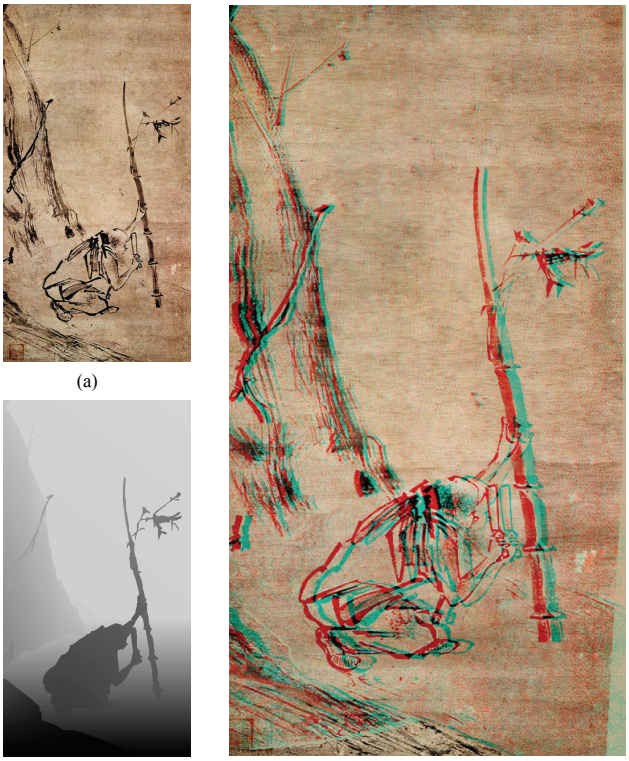


Figure 4: *Hangong ChunxiaoTu: (a) original map; (b) depth map; (c) anaglyph.*

Two TCPs of different styles are used to test the proposed technique. The first is *Hangong Chunxiao Tu* preserved in Taibei Gugong Palace Museum. It is one of the ten most famous ancient paintings in China and created by *Qiuying* in *Ming* dynasty. The whole painting, 37.2cm × 2038.5cm (height × width), depicts the life of ladies in the imperial palace, *Hangong*. Figure 4(a) shows a part of the painting. Figure 4(b) is its depth map. Here, the buildings are treated as a single layer since they are in the far background. The depth in each region varies in space. For example, the depth of the tree leaves ranges from that of the women touching the tree to that of the buildings behind them, simply by indicating a occupying region bridging the two in the image. Figure 4(c) presents the final stereoscopic image generated by the depth map.

The other painting is *Liuzu Zhuozhu Tu* (shown in Figure 5(a)), preserved in Tokyo National Museum. It is created by *Liangkai* in *Song* dynasty. The whole painting, 73cm × 31.8cm (height ×



(a) (b) (c)

Figure 5: *Liuzu Zhuozhu Tu: (a) original map; (b) depth map; (c) anaglyph.*

width), depicts a scene that *Liuzu*, the creator of Southern Zen Buddihism in China, is making firewood from bamboo. Behind him are mountains. Figure 5(b) and (c) present the computed depth map and anaglyph. The stereoscopic images should be viewed with red-cyan glasses (red on the left and cyan on the right). From the results, we can see that the method can deal with TCPs of different style, from yard to natural scenes. The results are convincing.

Compared with the manual approaches that could be used for TCPs [PSMethod1 2008] [PS2Method2 2010], the proposed method has the advantages as follows. 1) The developed tool depends on fundamental rules of the structural composition. 2) The tool is more convenient for beginners to use than the complicated software used in the the manual methods. To compute the ground depth, only seconds of interaction are required (drawing line cues). Besides, only two polygons are required for depth computation of each non-ground region, which can be finished in seconds. 3) The polygon interpolation model can generate smoothly varying depth of a layer in vertical and horizontal directions. Oppositely, manual methods assume a constant depth map of an object [PSMethod1 2008] or brush out a convex or concave boundary pixel by pixel [PS2Method2 2010]. 4) Pre-segmentation is required in most interactive methods [PS2Method2 2010] [Gimpel3D 2011]. In the proposed method, for objects with concise boundaries, e.g. buildings, no pre-segmentation is needed as described in Section 3.2.

5 Conclusion

This paper presented a simple technique for converting traditional Chinese paintings to stereo. The task was decomposed into depth computation of ground and non-ground pixels. A user interface was provided to indicate cues for the computation. The design of algorithms and user interfaces were derived from the principles of moving focus based spatial composition used in TCPs. The generated stereoscopic images are convincing.

Two issues need further effort. First, in new views generated through translation by disparities, some originally invisible and not existing parts are visible. In the experiments, these blanks are filled with the colors of the canvases (Section 3.3). In the future, sophisticated methods should be developed for more convincing results. Second, the polygon interpolation can generate varying layer depth in vertical and horizontal directions (Section 3.2) but local convex/concave depth. In the future, we will try to improve the technique to be capable of, for example, making the arm of *Liuzu* in Figure 5 cylindrical convex.

Acknowledgements

This work is supported by ETRI project no. EA2011489 and NSFC project no. 61003105.

References

FONG, W. C. 2003. Why Chinese painting is history. *The Art Bulletin 85*, 258–280.

GIMPEL3D. 2011. Creating stereoscopic images using gimpel3d. *http://gimpel3d.com/*.

GUTTMANN, M., WOLF, L., AND COHEN-OR, D. 2009. Semi-automatic stereo extraction from video footage. In *Proceedings of International Conference on Computer Vision*, 136–142.

JOBSON, D. J., RAHMAN, Z., AND WOODELL, G. A. 2010. Single image depth estimation from predicted semantic labels. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1253–1260.

KIM, D., MIN, D., AND SOHN, K. 2008. A stereoscopic video generation method using stereoscopic display characterization and motion analysis. *IEEE Transactions on Broadcasting 54*, 2, 188–197.

NEDOVIČ, V., MIN, D., AND SOHN, K. 2010. 3-D depth reconstruction from a single still image. *International Journal of Computer Vision 76*, 1, 53–69.

NEDOVIČ, V., MIN, D., AND SOHN, K. 2010. Stages as models of scene geometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence 32*, 9, 1673–1687.

PS2METHOD2. 2010. Creating stereoscopic images using photoshop. *http://3dvision-blog.com/converting-a-2d-image-into-a-stereoscopic-3d-image-with-photoshop/*.

PSMETHOD1. 2008. Creating stereoscopic images using photoshop. *http://www.3dphoto.net/forum/index.php?topic=1283.0*.

ROWLEY, G. 1947. *Principles of Chinese Painting*. Princeton University Press.

TORRALBA, A., AND OLIVA, A. 2002. Depth estimation from image structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence 24*, 9, 1226–1238.

YANG, X., BARNHART, R., NIE, C., CAHILL, J., LANG, S., AND WU, H. 1997. *Three Thousand years of Chinese Painting(The Culture & Civilization of China)*. Yale University Press.

ZHANG, G., JIA, J., TSIN WONG, T., AND BAO, H. 2009. Consistent depth maps recovery from a video sequence. *Transactions on Pattern Analysis and Machine Intelligence 31*, 6, 974–988.